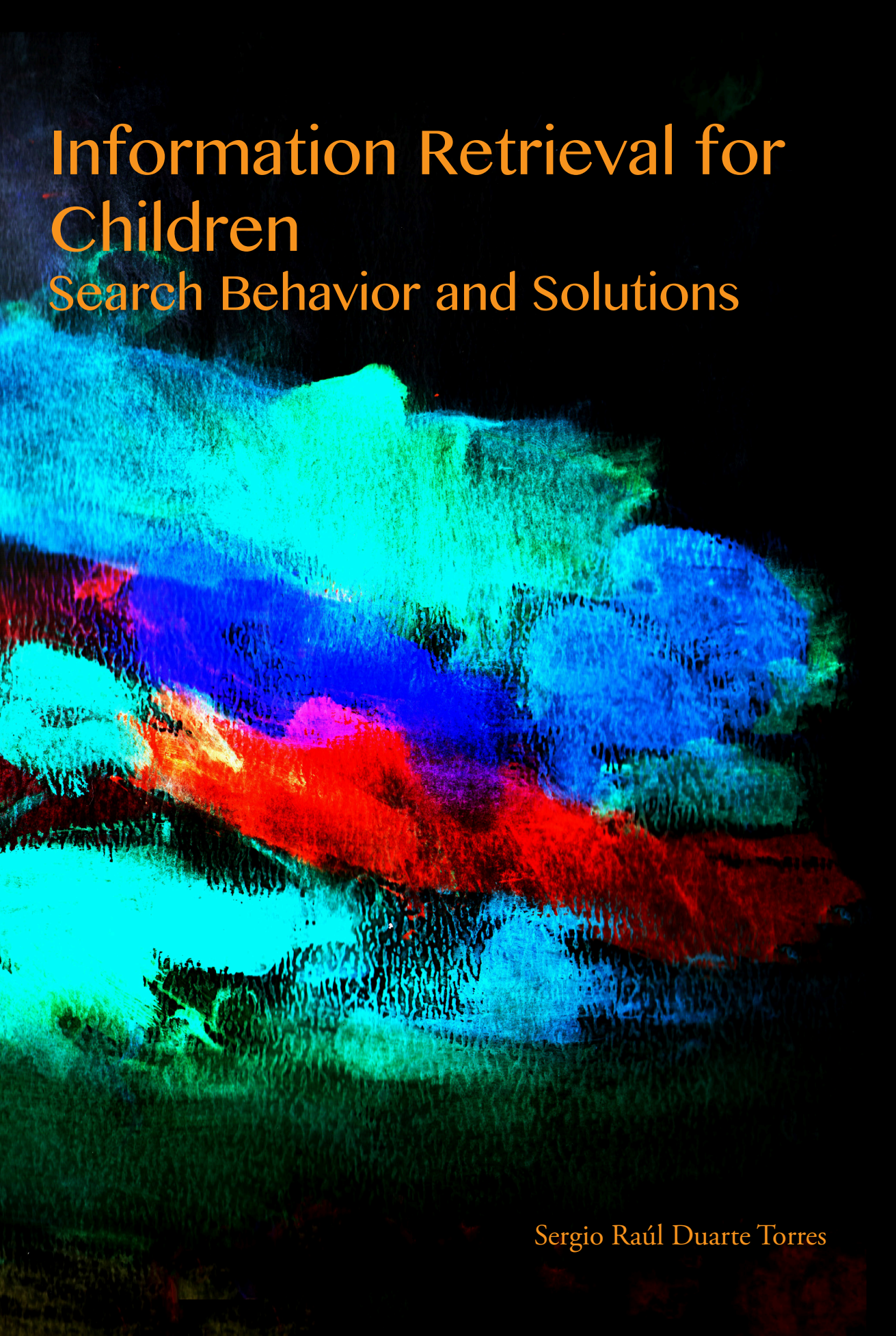
Information Retrieval for Children

Search Behavior and Solutions



Sergio Raúl Duarte Torres

Information Retrieval for Children: Search Behavior and Solutions

Sergio Raúl Duarte Torres

Graduation committee:

Chairman and Secretary	
Prof. dr. P. M. G. Apers	University of Twente, NL
Promoters	
Prof. dr. P. M. G. Apers	University of Twente, NL
Prof. dr. T. W. C. Huibers	University of Twente, NL
Assistant promoter	
Dr. ir. D. Hiemstra	University of Twente, NL
Members	
Prof. dr. Alan Smeaton	Dublin City University, Ireland
Prof. dr. Arjen de Vries	TU Delft / CWI, NL
Prof. dr. Franciska de Jong	University of Twente, NL
Prof. dr. Dirk Heylen	University of Twente, NL
Dr. Jaime Arguello	University of North Carolina at Chapel Hill, USA

CTIT Ph.D. Thesis Series No. 14-295
Centre for Telematics and Information Technology
University of Twente
P.O. Box 217, 7500 AE Enschede, NL

SIKS Dissertation Series No. 2014-03
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

ISBN	978-90-365-3618-9
ISSN	1381-3617
DOI	10.3990./1.9789036536189

Cover Design	Avril Follega/Sergio Duarte
Printed by	Gildeprint Drukkerijen

INFORMATION RETRIEVAL FOR CHILDREN: SEARCH BEHAVIOR AND SOLUTIONS

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
Prof. dr. H. Brinksma,
on account of the decision of the graduation committee
to be publicly defended
on Friday, 14th of February, 2014 at 16.45

by

Sergio Raúl Duarte Torres

born on December 19, 1983
in Bogotá, Colombia

This dissertation is approved by:

Prof. dr. P. M. G. Apers (promotor)
Prof. dr. T. W. C. Huibers (promotor)
Dr. ir. Djoerd Hiemstra (assistent-promotor)

This thesis is dedicated to the loving memory of my beloved father. May his memory forever be a source of inspiration and blessings. Thanks for everything.

Acknowledgements

I have finally completed my doctoral thesis after something more than four years. It has been a long way with ups and downs, and foremost, with plenty of learning.

In these pages I want to express my gratitude to the persons that supported me, gave me a hand, and in general to the persons that provided me their company in this journey. All of them played and important role in the completion of this thesis and a very important role in making fun the time I spent developing this work.

First at all, I want to thank my parents who did not only raise me and take care of me, but also always make their best efforts to make my dreams possible. I deeply appreciate the advices and company of my sister during the difficult moments. I am also very thankful to Avril for sharing her playful attitude and joy towards life, and of course, for designing the cover of this thesis.

I would like to express my deepest gratitude and appreciation to my promoter Peter Apers and my co-promoter Theo Huibers, who gave me the opportunity to carry out the research presented in this thesis, and more importantly, for giving me the chance of becoming a doctor.

Foremost, I want to thank my daily supervisor, Djoerd Hiemstra, for his continuous guidance, motivation, patience and vast knowledge. I am certain I would have not reached this point without his advises, and I am certain that I am better professional and person after his mentoring.

Besides my supervisor, I would like to thank the other members of the committee: Dr. Alan Smeaton, Dr. Jaime Arguello, Dr. Arjen de Vries, Dr. Franciska de Jong and Dr Dirk Heylen, for their generous time and good will through the preparation and review of this thesis. I feel very honored to have such a distinguished and tough committee.

I am also very grateful to all the members of the DB group. I would particularly like to thank Ida and Suse, who helped me in a myriad ways, from giving me tips about the Dutch life to let my PhD going all these months. I also really appreciate Jan's support with the IT issues. I greatly enjoyed the conversations about his trips, which were very encouraging.

Many thanks to Riham who helped me to get along with the Dutch life during my first months in the Netherlands; to Rezwan for his company, especially during the first months of my staying in the UT; to Juan for all those conversations about technical and random stuff that came out every time I visited his office; to Zhemin and Mohammad for the nice conversations, and to Christoph, Mena, Brend, Victor, Iwe, Almer and Lei for their company during the daily life in the group.

I own a very important debt to all the members of the PuppyIR project, from whom I did not only receive some tasks, but also plenty of feedback and enlightening moments. Particularly, I highly appreciate the constructive comments and warm encouragement of Arjen de Vries, Franciska de Jong, Ian Ruthven, Leif Azzopardi, Andreas Lingnau and Marie-Francine Moens. Many of those comments shape the structure of my thesis.

I want to extend my appreciation to Carsten Eickhoff, Ric Glassey, Frans Van der Sluis and Tamara Polajnar for giving me the chance of cooperating with them in some of the endeavors of the project and the publications that were the results of that effort. I also highly appreciate the guidance and tips provided by Hanna Jockmann-Mannak these months.

I woud like to show my gratitude to Pavel, as a former member of the DB group and the PuppyIR project, for his mentoring at the beginning of my PhD, for his ideas, suggestions and most of all for his encouragement.

An important part of my research was carried out at Yahoo! Labs in Barcelona. I want to thank all the members of the lab. Particularly, I am very thankful to Ricardo Baeza-Yates for facilitating my stay in the labs and I highly appreciate the discussions and feedback offered by Mounia Lalmas, first in the PuppyIR meetings and then in the labs.

I owe sincere and earnest thankfulness to Ingmar, for his assistance and mentoring. I really appreciate his patience and tenacity, which helped me make the most of my time in the lab. I learned a big deal from him.

I also want to thank Hugues Bouchard for making my stay in Barcelona lively and showing me all those funky places in Gótico. Thanks a lot to Marek, Luca, Eraldo, Bart, Ruth, Giorgos, Jannete and Michele for the fussball matches and for hanging out in the city. Special thanks to Erik for his company and especially for his geeky tips and jokes during our stay.

I have many people to thank beyond the scope of the campus and the research world. Firstly, I want to say thank you to the Piepke community, a.k.a my housemates. I am very thankful to Robin, as a colleague and even

more as a housemate, I really appreciate his advices and his friendship. I also want to thank Marga for her kindness and everlasting tolerance of all my Latin habits. I also appreciate the time as housemates with Danielle, and more recently with Hans, Diego and Vincent.

In these long months there were two persons that were always there when I need them. I will be always in depth debt with Maral for keeping my sanity through all our conversations and good coffee moments. I am also very glad to have met Juan Jimenez from the very early stages of my PhD, he has been a great friend and a person I can always count on.

I am also glad to be part of the Latin move of Enschede, including my year as a board member of La Voz. It was very rewarding to be part of the board along with Oscar, Teo, Diruji, John, Juan Carlos and Daniela.

During my stay in Enschede I met wonderful people. I feel blessed for meeting Andrea, Andreita, Andreea and Adriana, for their friendship and dancing lessons. It was a pleasure to attend the gigs of Chilangos and Rimon, so thanks a lot to David, Oscar, Jorge, Diego and Gerard. It was great to count with Abraham, Cesar, Norma and Israel at lunchtime, especially because everybody else likes to have lunch too early. I thank Eduardo and Nayeli, for their company during my first months in Enschede and for encouraging me to play fútbol, although it did not last long, it was great fun. I also enjoyed greatly the meetings with Vicky, Lorenzo, Jorge and Maite in all the parties and dinners.

Talking about parties, I had some very nice moments in places like the Molly Malone and, of course, the Paddy's, particularly with Desmond, Juan Pablo, Nico, Ignacio, Marina, Pablo, Jenny, Arturo (both), Javier, and many of the persons I have mentioned so far.

For the times I was back home, I have to thank los compas, for all the catching up, the fútbol games, the drinking and their immense encouragement. The company of Goyes, Barrantes, Juan Manuel, Juan Carlos, Lina, Lis, Raul and Camilo were refreshing and completely necessary to recharge my energies prior to every year in the Netherlands. Without those moments I am sure I would have taken much longer to finish this thesis.

Among my compas, I am always in debt with David, who has filled me with motivation for so many years and who has been there all the time during my PhD, even in the long distance, his friendship has always been invaluable. I am also very grateful to Jairo. I really treasure his unconditional friendship that started from the early days at school; friendship

and good memories that I am sure will remain.

I am also want to thank my lovely Anouk, for being a constant source of motivation and encouragement. I also appreciate her patience, especially in my cranky moments.

I apologize for all the persons I have not mentioned, I am sure some names escape my mind now, which does not mean that they did not play or keep playing an important role in my life. Many thanks to them as well.

Abstract

Nowadays, children of very young ages and teenagers use the Internet extensively for entertainment and educational purposes. The number of active young users in the Internet is increasing everyday as the Internet is more accessible at home, schools and even on a mobile basis through cellphones and tablets.

The most popular search engines are designed for adults and they do not provide customize tools for young users. Given that young and adult users have different interests and search strategies, research aimed at understanding the activities that young users carried out on the Internet, the way the search for information, and the difficulties that they encounter with state-of-the-art search engines, are urgently needed. The first contribution of this thesis addresses these research aims by providing a characterization, on a large scale, of the search behavior of young users. The problems they face when they search for information on the web, the topics they searched and the online activities that motivate search were explored in detail and contrasted against the search behavior of adult users. The results presented in this thesis have important implications for the development of search tools for young users and for the design of educational literacy.

Two central problems were identified in the search process of young users: *(1)* difficulty representing the information needs with keyword queries, and *(2)* difficulty exploring the list of results.

We found that focused queries are often required to access high quality content for young user with modern search engines. However, young users were found to submit queries that lack the specificity needed to retrieve content that is suitable for them, which leads to frustration during the search process. This observation motivates the second contribution of this thesis. We propose novel query recommendation methods to improve the chances of young users to find content that is suitable and on topic. Concretely, we present an effective biased random walk based on information gain metrics. This method is combined with topical and specialized features designed for the information domain of young users. We show

that our query suggestions outperform by a larger margin not only related query recommendation methods but also the query suggestions offered by the search services available today.

In respect to the second difficulty, it was found that young users have a strong click bias, in which results ranked at the bottom of the result list are rarely clicked. This behavior greatly hampers their navigational skills and exploration of results. It also reduces the chances of young users to find suitable information, since appropriate content for this audience is ranked, on average, at lower positions in the result list in comparison to the content aimed at the average web user.

The third contribution of this thesis aims at helping young users to improve their chances to find appropriate content and to ease the exploration of results. For this purpose, we envisage an aggregated search system in which parents, teachers and young users add search services with content of interests for young audiences. We propose a test collection with a wide number of verticals with moderated content, a carefully selected set of search queries and vertical relevant judgments. We also provide novel methods of vertical selection in this information domain based on social media and based on the estimation of the amount of content that is appropriate for young users in each vertical. We show that our methods outperform state-of-the-art vertical selection methods in this information domain.

We also show in a case study with children aged 9 to 10 years old that result pages derived from the collection proposed are preferred over the result pages provided by modern search engines. We provide evidence showing that the interaction and exploration of results are improved with result pages built using this collection, even if the users of this case study were unaware between the differences between the types of pages displayed to them.

This thesis is concluded by providing concrete follow-up research directions and by suggesting other information domains that can potentially benefit from the methods proposed in the thesis.

Contents

List of Figures

Chapter 1

Introduction

The fraction of children using the Web and the amount of time they spend online has increased significantly in past years. A case study, carried out in 2008, involving up to 2,500 in-home interviews with children and their parents in the UK, reported that 63% of users aged 5 to 7 and 76% aged 8 to 11 years old, use the Internet at home [Child Trends Data Bank, 2013]. In the US, 32.4 million children under the age of 18 years old were active users of the Internet in the same year, accounting for up to 19% of the online population. Similar trends have been reported in other developed countries [Ofcom, 2010]. More recent studies carried out in the European Union have pointed out that not only has access to the Internet continued to increase among the young population but also the amount of time they spend online. Livingstone et al. [2011], using a detailed survey carried out from 2009 to 2011 of European children and their parents in 25 countries, reported that users from 9 to 16 year old spend on average 88 minutes per day online. They also found that 33% of these users go online via mobile phones, 87% at home, and 49% at home from their bedroom. Even higher Internet access percentages have been reported in the last years for this segment of users in the US.

Madden et al. [2013] found, through a survey conducted in 2012 with 802 parents and their teenagers aged 12 to 17 years old, that 95% of the teenagers use the Internet regularly, 78% own a cell phone and around 47% own a smart-phone, which is a prominent means of accessing the Internet. It has also been shown that children are often trusted to search the Internet on their own, 68% and 84% of children aged 5-7 and 8-15 in the UK, respectively [Ofcom, 2010]. Undoubtedly, the access and use of the Internet by children will keep increasing in these and other regions of the world in the coming years.

Most of the current Information Retrieval (IR) systems are designed for adults and previous case studies have shown that the information needs and search approaches of children and adults differ substantially [Bilal and Watson, 1998; Broch, 2000; Druin et al., 2009, 2010; Nahl and Harada, 1996].

For instance, children have been found to be less focused during the search process. They were often observed following a non-linear navigational style, in which resources previously explored are reactivated [Bilal, 2002] Lack of logical progression in the exploration of results was also observed. This behavior exposes the disorientation children experience when they search the web and the difficulties they face in deciding which information is relevant [Bilal, 2001]. Difficulty constructing meaning from the results has also been reported in children, especially in case of complex information needs that required pieces of information from different sources [Bilal, 2001]. These studies have been very useful in identifying some of the difficulties children experienced when asked to solve information tasks on the Internet.

Nonetheless, these studies are highly obtrusive and they only consider a limited number of users, making it hard to extrapolate the results to a larger scale. Obtrusive studies refer to the gathering of measurements when the subjects are aware that they are being observed. This awareness leads the subjects involved in the study to change their behavior and responses, which can greatly affect the validity of the data gathered during the experimental process [Webb et al., 1981]. In the particular case of child users, it has been acknowledged that children (and even adults) often lack the objectivity to describe accurately their behavior and task outcomes, especially when situations involve adverse factors [Pettersson et al., 2004].

1.1 Search and Browsing Behaviour of Children on a Large Scale

The first aim of this thesis is to characterize the way children up to 12 years old and teenagers from 13 to 18 years old search the Web, and to measure the struggle of these users to find information on the Web using well established query log metrics and novel metrics, especially tailored to this user segment. In Chapters 2 and 3 we will break down the group of children and teenagers into more detailed age groups based on the characteristics of the data analyzed.

Similarly, little is known about the activities that young users engage in outside the search box. Recently, these activities, referred to in this thesis as *browsing* activities, have received the attention of the research community for the case of the average web user (disregarding of the users' age) [Cheng et al., 2010; Goel et al., 2012; Kumar and Tomkins, 2010]. The second research aim of this thesis is to characterize the *browsing* activities that young users engage in on the Internet and to identify the type of browsing activities that are more likely to trigger search in state-of-the-art search engines. An integral understanding of search behavior is obtained by analyzing the search process within a search engine along with the activities that lead to search queries being submitted.

Our approach differs from previous attempts to understand the search and browsing behavior of young users in that we quantify the search characteristics based on the aggregated results of thousands of users across a broad age range, unobtrusively, which makes our observations more representative on a web-scale. Moreover, we explore the browsing activities that motivate searches made by young users.

The first research aim is addressed through the study of two large scale query log sets extracted from the AOL and Yahoo! search engines. In the case of the AOL search logs, a set of queries which lead to trusted resources for young users is identified. We employ this set and their search sessions to analyse the differences in the query and session characteristics of users searching for information suitable for children between 10 to 12 years old and content for teenagers (users between 13 to 15 years old) against users searching for general purpose content. We will show notable differences in the search behavior of these users and the topics searched. Note that by using the AOL search logs we are unable to ensure that the queries extracted were actually submitted by young users, however we can be assured that the users were interested in content clearly orientated for this segment of users. Nonetheless, through this set of queries we are not only able to characterize the search behavior of users searching for content for young users but also able to quantify the difficulties of reaching high quality content for them. Chapter 2 presents details of how the extraction of the search sessions and their findings are derived from this analysis. Concretely the following research questions are addressed:

R.Q-1.1: What are the differences in search behavior of users targeting content for young users in respect to the average web user?

We are also interested in verifying that the queries extracted from clicks on trusted resources for young users are representative of the topics searched by the queries submitted by these types of users:

R.Q-1.2: Can we identify a representative distribution of topics of interest to young users in the Web through a set of queries aiming at content for them?

For this purpose, we compare qualitatively and quantitatively the distribution of topics obtained from the AOL and Yahoo! search logs. For the latter we extract only search activity from users with a registered profile. In this way we are able identify the age of the users submitting the queries. The analysis of these logs differs from the one carried out with the AOL search logs in that we are able to estimate the age of the users, which allows us to apply a user-centric approach in the analysis of the search sessions.

Thus, the results derived from the AOL log are representative of users clicking on high quality content for young users while the results derived from the Yahoo! search

logs are representative of the average search behavior of users of specific age ranges. The queries from the Yahoo! search logs are grouped using fine-grained intervals from users aged from 6 to 18 years old (i.e. 6-7, 8-9, 10-12, 13-15, 16-18). We hypothesize that the large and complex volume of information to which young users are exposed leads to ill-defined searches and to disorientation during the search process.

R.Q-2.1: Do young users struggle to find information with a large scale search engine, and how is this struggle reflected in their search behavior from a query log perspective?

We quantify their search struggle based on query metrics (e.g. fraction of queries posed in natural language), session metrics (e.g. fraction of abandoned sessions) and click activity (e.g. fraction of ad clicks). We will show that these metrics clearly demonstrate an increased level of confusion and unsuccessful search sessions among young users.

A comparison between young and adult users is key to identifying the current deficiencies of state-of-the-art search engines in supporting the search process of young users and satisfying their information needs. For this purpose, an analogous analysis was carried out for adult users over 18 years old.

R.Q-2.2: Does the search behavior and search difficulties of children, teenagers and adults differ in a large scale search engine (Yahoo! Search)?

We also hypothesize that the development stages of children and teenagers are reflected in their web searches and that this development can be traced through the search logs:

R.Q-2.3: Can we retrace stages of children and teenagers development, in terms of the topics they are interested in, through their queries and the characteristics of these queries?

The tracing of specific aspects of human development has a wide variety of applications not only for search engines designers but also for professionals in child care and related areas. For this research question we focus on the changes in the users interests (e.g. distribution of topics searched), language development (e.g. readability of the content accessed) and cognitive development (e.g. sentiment expressed in the queries) among children, teenagers and adults. We will show that our findings can be exploited to lead to a more relevant selection of information services for young users. Chapter 3 describes in further detail each one of these research questions and discusses our findings.

The second research aim is addressed by analysing a large sample from the Yahoo! toolbar logs, in which all the urls entered by the user to the Web browser are captured,

including the activities that do not occur within the standard search engine. Note that with AOL and Yahoo! search logs, only the events within the standard Web search are studied. The Yahoo! toolbar logs allow us to explore the usage of non-web search functionality, such as *Image* and *Video* search. Search on these two services will be referred to as *multimedia* search and their usage will also be explored across the age ranges defined in the analysis with the Yahoo! search logs.

The aim of this research can be subdivided into two steps: identifying users browsing activities (and multimedia search activity) according the age of the user, and measuring the likelihood of each one of these browsing activities to trigger a search:

R.Q-3.1: What activities are carried out by young users on the web browser besides web searches?, How prominent is browsing for each age range? At what ages are multimedia searches preferred?

To address *R.Q-3.1*, a broad set of browsing activities are employed to classify the frequency in which users of different ages engage in each one of these activities, which range from social activities (e.g. Facebook) to knowledge oriented browsing (e.g. Wikipedia). As it was mentioned before, an integral view of search behavior involve the understanding of how the browsing activities trigger searches in young and adult users:

R.Q-3.2: Which types of search and browsing activities are more likely to trigger searching the web and multimedia search engines in the case of young users? Do these triggers differ from those observed in adults users?

To address *R.Q-3.2* we quantify the proportion of browsing and search activity in the toolbar sessions and we estimate the likelihoods of carrying out a search on the web search engine and multimedia search engines (i.e. videos and images) given that the previous event is another search event or browsing event. We will show that children tend to engage their activities on the Internet through a search engine more often than adults and that multimedia search is preferred within specific age ranges. The results of this analysis and recommendations for future work are presented in Chapter 4.

1.2 Aiding Young Users to Search the Web

Two central problems that arise during the search process of young users with state-of-the-art search engines are: *(1)* difficulty representing the information need with keyword queries, and *(2)* difficulty exploring the list of results. These difficulties

are described in further detail throughout chapters 2, 3 and 4, nonetheless they are briefly described in this section to introduce the two solutions explored in this thesis.

We observed in the two search log analyses that elaborated queries are often required to access high quality content for young users. Elaborated queries refer to queries that are longer than those submitted by the average web user and that contain keywords to focus on the retrieval of content that is suitable for young users. However, these users submit queries that lack the specificity to retrieve content that is suitable for them, which leads to frustration during the search process.

We found a greater usage of natural language in younger users, which has also been observed in previous case studies with small groups of children [Bilal, 2001, 2002]. This behavior, along with the significant difference in the vocabulary size observed between queries submitted by young and adult users, mark the importance and urgency of providing adequate query assistance tools for this audience, especially considering that the query formulation represent the first step in the seeking information process.

In regard to the second difficulty, it was found that young users have a higher click bias than adults which lead to a lower volume of clicks on lower ranked results, behavior that greatly hampers their navigational skills and exploration of results. This behavior is particularly problematic given that the content appropriate for this audience was observed to be ranked lower in the results page (details can be found in Section 2.6.7). Foss et al. [2012] also reported that certain groups of children (*domain searchers*) only search a limited number of websites or specific domains (i.e, gaming), behavior that lead to search breakdowns when children searched for content in different domains and when unseen web resources needed to be explored. The lack of focused queries and prominent click bias on top ranked positions exposes young users to content that is not on target and, in some cases, that can be harmful to them, since current search engines provide information for all kinds of public.

In this thesis we explore two solutions to address the search problems mentioned above. These solutions target the main recommendations derived from the research question described above: *(1)* query recommendation and *(2)* resource selection for young users. In the following paragraphs we will describe each solution:

1.2.1 Query recommendation for young users

We address the first search difficulty by proposing a novel query recommendation method based on a biased random walk that emphasizes the query aspects related to content of interest for young users. The method utilizes tags from social media to suggest queries related to young users topics. The evaluation is carried out using a large scale query log sample of queries submitted by young users, classified in fine-grained age ranges. The query suggestions attempt to close the vocabulary gap of

young users, and more particularly they provide focused queries targeting content adequate for this public.

The evaluation is carried out using a large query log sample of queries submitted by young users that lead to successful clicks. We show that our method outperforms, by a large margin, the query suggestions of modern search engines and state-of-the-art query suggestions based on random walks:

R.Q-5.1: To what extent does a random walk, biased by using information gain metrics, improve the effectiveness of the query recommendations for young users over traditional biased and unbiased random walks?

We further improve the quality of ranking by combining the score of the random walk with topical and language modelling features to emphasize further those query suggestions that represent topics and information aspects suitable for young users. The evaluation of this approach is used to address the following research question:

R.Q-5.2: Can we improve the quality of the ranking of query recommendations by combining the random walk score with features based on language models and topical knowledge?

A detailed description of the two methods and an extensive evaluation are presented in Chapter 5.

1.2.2 Resource selection for young users

We envisage an information retrieval system that builds on the aggregated search paradigm to address the search difficulties mentioned in the previous section in a holistic fashion [Duarte Torres, 2011; Murdock and Lalmas, 2008]. Aggregated search refers to the selection of results from diverse search services or search engines and the presentation of these results on an single result page by organizing them in a coherent way, beyond the classic result list provided by modern search engines [Gyllstrom and Moens; Kopliku, 2009]. These search services are often referred to as verticals, which are defined as domain specific collections, (e.g. entertainment, shopping, news, recipes) or collections of specialized types or genres (e.g. videos, images, songs).

The system envisaged integrates heterogeneous content from verticals which are not fully accessible to the system (third party verticals). In particular we are interested in verticals that contain high quality information for children from 8 to 12 years old. In this system, parents, teachers and other specialists in child care would be allowed to add resources for children. For instance, they could add a vertical dedicated to coloring pages: `http://ivyjoy.com/colouring/search.html`, which only returns sheets of paper to be colored and that are suitable for children, or a vertical dedicated to search only videos: `www.youtube.com`, in this case the vertical

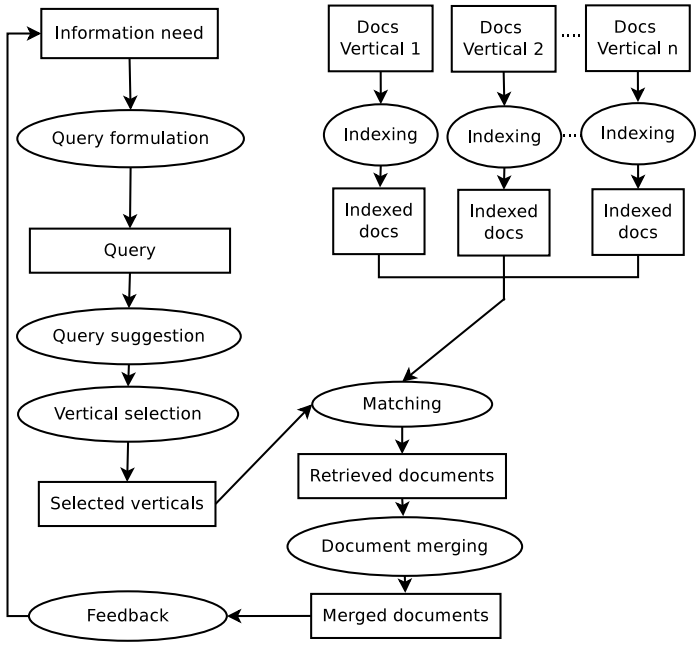


Figure 1.1: Aggregated search process

provides content for all kinds of public segments. We believe that an aggregated search system is a better solution for searching content on the web for children than simply crawling and indexing websites because *(i)* is more scalable, and *(ii)* we can leverage and exploit the knowledge of parents.

Figure 1.1 shows the information retrieval process for the case of Aggregated search [Croft, 1995; Murdock and Lalmas, 2008]. The search starts with the user formulating a query. Afterwards the system generates query suggestions which are displayed to the user. Note that the query recommendation process is not an exclusive step in the aggregated search paradigm, since it also uses part of Croft's information retrieval process [Croft, 1995], however this step is still crucial. A step beyond recommending queries to the user is to recommend a set of specialized search engines or collections in addition to the standard web set of results. Recommending specialized collections would greatly help young users focus their search on content that is oriented to their information needs and that is more appropriate in terms of the quality of the content and the media genre. This step is referred to in Figure 1.1 as *vertical selection*. For instance, consider the query, *math coloring puzzles*. A state-of-the-art search engine provides a list of web results where coloring sheets can be found after exploring the urls. However, children explore less results than adult users, which means that they are less likely to get to the targeted content. On the other hand, recommending a search engine specialized in coloring pages and displaying the images on the result page would highly reduce the burden of having to explore the web results. In this way, young users can find their way to the information faster

and in a direct manner, which is one of the main difficulties young users face during the search process, as pointed out in the previous section. This solution also aims to improve the accessibility to genre specific verticals. Results from visual verticals are particularly important for certain types of search tasks in which it is easier to find information using visual or genre specific content. For instance answers for the query list of american presidents can be found 28% faster using image search instead of the standard web search[1].

Chapter 4 will show that providing rich media from different genre verticals can greatly improve the web experience of young users, since users between 10 to 19 years old are around 2.4 times more likely to submit a query on a multimedia vertical than adult users after browsing content on the Internet. Improving the accessibility to results from non-standard verticals is particularly important for users below 10 years old since they have a harder time finding this type of content, as will be shown in Chapter 4. Similar observations have been drawn in previous research. Bilal [2001]; Druin et al. [2010] reported that users in these age ranges have difficulties in identifying the tabs and hyper-links of the non-web verticals.

In Chapter 6 we explored the usage of vertical selection methods in the specific domain of topics for children between 8 and 12 years old. A test collection with an extensive set of queries, verticals and relevant judgement is described. The selection of queries and verticals used to build the collection are based on the findings shown in Chapter 2 and 3 respectively. In the latter, we observed marked differences in the distribution of topics targeted by queries of children, teenagers and adults. The topical division between users of different ages suggests that personalizing the results from a selection of verticals according the age of the user is an effective strategy to focus the search on topics that are more likely to be of interest to the user.

Two methods are explored for the vertical selection problem in the domain of content for young users. In the first method, the global and domain specific sizes of the verticals are estimated. These estimations are used together with state-of-the-art methods of vertical selection to improve their performance. In the second method a novel vertical and query representation was introduced based on tags from social media. We show that the use of tags from social media lead to significant performance gain.

The evaluation of both methods is used to address the following research questions:

R.Q-6.1: To what extent can we improve state-of-the-art techniques of vertical selection through the estimation of the content available in the verticals for users between 8 to 12 years old?

R.Q-6.2: What is the benefit of using tags from social media to represent the query and the verticals in the problem of vertical selection?

[1] `http://www.nytimes.com/2009/12/26/technology/internet/26kidsearch.html`

Both approaches are contrasted in isolation against two well-established methods of vertical selection. Additionally, we will show that the combination of the methods (by weighting the scores) lead to better performance under specific experimental settings. It is important to mention that the evaluation of *R.Q-6.1* and *R.Q-6.2* is carried out by using vertical relevance judgement submitted by adult users. Two limitations can be pointed out in regard to the vertical selection study: *(i)* the benefits of presenting results from the verticals of our collection to actual young users is unknown, and *(ii)* even though it is reasonable to assume that adults are able to identify content suitable and relevant for children, it is unknown if these vertical preferences differ from the preferences of young users. These limitations are rewritten in research question as follows:

R.Q-6.3: Do users aged 8 to 12 years old explore more pages with blended results from the verticals of our collection than pages retrieved from a state-of-the-art search engine? In which type of result pages do they agree more in terms of the content clicked?

R.Q-6.4: Which verticals are preferred by children aged 8 to 12 years old given an heterogeneous set of topics, and how do these vertical preferences differ from the preferences of adult users?

A case study was carried out with a group of school children aged 9 to 10 years old and a group of adult users. The group of adults were addressed through crowd-sourcing. In the two case studies, users were asked to engage in a game designed to evaluate the interaction and exploration of results in pages blending results from the verticals of our collection and pages with results from state-of-the-art search engines. The game consists of selecting results for open information tasks. The more users clicked on a given result, the more points were awarded for a click on the result. *R.Q-6.3* is addressed by comparing the number of clicks, points awarded, click user agreement and likelihood of clicking on vertical results in the different types of aggregated pages shown to the users. *R.Q-6.4* is addressed by comparing the vertical preferences of both group of users and by measuring their agreement on the page types evaluated. Special attention is given to the specific vertical disagreement between the two type of users.

1.3 Thesis outline

In Chapter 2 the query log analysis of queries aiming at content for young users based on the AOL search logs is presented. Chapter 2 is based on Duarte Torres et al. [2010a,b]. Chapter 3 presents an extensive analysis of queries from registered users, with reported age, from the Yahoo! search logs, focus of search difficulty and

differences between fine-grained age groups is also carried out in this chapter. Chapter 3 is based on Duarte Torres and Weber [2011]; Duarte Torres et al. [2014a]. Chapter 4 presents a large scale analysis of browsing behavior through the Yahoo! toolbar logs. The emphasis of this chapter is on browsing activities that trigger or motivate search in users of different age ranges. Chapter 4 is based on Duarte Torres et al. [2014a]. Chapter 5 explore methods of query recommendation for young users based on biased random walks and topical features. This Chapter is based on Duarte Torres et al. [2012, 2014b]. In the first part of Chapter 6 presents a description of the test collection built for the problem of vertical selection in the domain of young users (particularly children) and present two novel mechanism to perform this task in the targeted domain. In the second part of Chapter 6 is described the game used to engage the group of school children and adults and the results of the case studies, which provide clear evidence of the benefits of blending results from the verticals of our collection to children. Chapter 5 is based on Duarte Torres et al. [2013]. Chapter 7 summarizes the findings of this thesis and presents ideas and direction for follow up research.

Chapter 2

Search behavior of users when targeting content for young users

This chapter is based on Duarte Torres et al. [2010a,b].

2.1 Introduction

Given the small amount of content carefully designed for this audience and the lack of specialized search engines dedicated to help children to find appropriate content on the Web, there is an increasing need for research aimed at understanding the search behavior and difficulties that these users experience. This is a highly important matter that requires research considering that children's information needs [Walter, 1994], search approaches and cognitive skills differ from those of adults [Kuhlthau, 1991a].

Query logs represent valuable sources of information to understand the search process and to improve search engine systems. For instance, query logs have been widely exploited in the literature to study user's behavior/interaction with IR systems, to classify queries [Chien and Immorlica, 2005], infer search intent [Baeza-Yates et al., 2006; Broder, 2002], to generate user profiles [Baeza-Yates et al., 2006], to produce query suggestions [Boldi et al., 2008], among others.

Recall that the two main advantages of query logs usage for the study of the search process are *(i)* the large scale scope of the analysis, and *(ii)* the unobtrusive nature in the collecting of data. A large scale analysis provides a representative overview of the type of information needs and topic interests that young users have on the Internet, which can not be obtained from small case studies.

In this chapter we explore the AOL query log [Pass et al., 2006] to compare queries and sessions used to retrieve information for young users and to retrieve general purpose information. The following research questions summarize the aims

of this chapter:

- **R.Q-1.1**: What are the differences in search behavior of users targeting content for young users in respect to the average web user?
- **R.Q-1.2**: Can we identify a representative distribution of topics of interest to young users in the Web through a set of queries aiming at content for them?

To address these research questions is necessary to identify queries aiming at content for children and teenagers. For this analysis we employ the Kids and Teens section of the *Dmoz* directory[1] as a gateway to identify queries employed to retrieve content for young users. The aim of this Dmoz section is to provide child friendly and safe content to cover the specific needs of people under the age of 18. We consider that using this directory to identify queries targeting content for young users is reasonable and realistic enough given that the content of this directory is frequently regulated and maintained by senior editorial staff, which guarantees that websites with harmful or unsuitable content for children are excluded.

Note that although it is not possible to establish if these queries were performed by young users, we are still able to study the characteristics of the queries and sessions when the underlying information need is related to content for young users. In the next chapter we analyze a large query log sample from the Yahoo! Search engine from users with reported aged. In the next chapter we also contrast the results obtained in this chapter using the AOL search logs and the Yahoo! Search logs.

In this and the coming chapters, we will refer to *children* as users aged up to 12 years old, and to teenagers as users aged 13 to 18 years old (young users refer to both groups). In this chapter we break down the *teenager* users section into *teens* and *mature teens*, which are aged 13 to 15 and 16 to 18 years old, respectively. This age classification is based on the *Dmoz* age tags, which can be used to distinguish between content suitable for kids up to 12 (kids), 15 (teens) and 18 years old (mature teens)[2]. In some cases we will refer to the *children* type of queries as *kids* queries to match the labelling provided by *Dmoz*. We will use a more detailed age segmentation in the next chapters.

In regard to the first research question, differences in the search behavior of users are accounted by looking at characteristics in the query space, search sessions and topic distribution of the log activity identified by using the *children* and *teenager* queries. Concretely, we will look into query length, natural language usage in queries, query intent (informational, navigational), query reformulation usage and session length. We will motivate the analysis of each one of these features in the following sections and we will contrast the results against previous findings of children (and teenager when applicable) search behavior.

[1] http://www.dmoz.org/Kids_and_Teens/
[2] http://www.dmoz.org/guidelines/kguidelines/

For the second research question, we will characterize the topics search by this set of queries using a cue word analysis based on clustering methods. Additionally, we map queries into the categories of the Yahoo! Directory [1] to identify differences in the topic distribution of queries aiming at children, teenager and the average web user queries.

It is important to mention that the two research questions posed in this chapter will be revisited in the next chapter, in which we analyzed a large search log containing queries with a high likelihood of been submitted by actual younger users. We will contrast the results obtained by both search logs.

This chapter is organized as follows: In Section 2.2 is described the most relevant related work on query log analysis and children search behavior. In Section 2.3 is described the research methodology employed in this study. Section 2.4 describes the data acquisition process. Section 2.5 describes the analysis and justifies our approach to compute metrics from the search logs. Section 2.6 presents the results obtained at a query level while Section 2.7 presents the results derived from the search sessions. In Section 2.8 is compared the results obtained with the Dmoz Kids and Teens section against the results of a different Dmoz category. We show that the results differ and not all the users identified through Dmoz behave in the same way. Finally in Section 2.9 conclusions are drawn and directions for future work are stated.

2.2 Related work

In this section is summarized the most relevant literature on children search behavior and query log analysis. For the latter we will make emphasis on analysis of search engine usage studies. We will make references to the findings of these studies when discussing our results from the AOL search logs.

2.2.1 Information seeking by children

The first studies attempting to characterize the search behavior of children have been carried out using non-internet systems, such as electronic libraries, CD-ROMs and OPACs (Online Public Access Catalogs). Solomon [1993] explored the search success of elementary school children when using an OPACs. The authors found that children were able to use the system effectively when engaging simple searches. However, they found that complex searches were hampered by the lack of mechanical skills of children. They pointed out that factors such as typing on the keyboard, spelling, limited vocabulary and reading expertise are skills that are not developed enough in children in order to use the OPAC system studied [Broch, 2000]. Borgman

[1] http://dir.yahoo.com

et al. [1995] found a similar behavior with high school children and a different OPAC. They also reported that these children had conceptual difficulties categorizing and browsing for searches that are more complex. Similarly, Neuman [1995] found from a survey including 25 digital library administrators, that the main problems children encountered during the search on digital libraries are the generation of keywords to construct the query, and the lack of effective search strategies.

Recent studies have explored the search behavior of children on the Internet with search engines. Nahl and Harada [1996] carried out a study with 191 high school students to determine their search effectiveness after they have received special training to search the Internet. Users were asked to solve specific information tasks on the Internet. They were assessed based on the information they collected. Nahl and Harada [1996] report that most of the students had difficulties understanding how the search query is constructed with boolean and default operators. In this study was also observed that the lack of adequate vocabulary and content knowledge led to difficulties in the search process.

Bilal and Watson [1998] conducted a case study with children from a 7th grade science class (children between 11 and 13 years old) to determine how this group of users solve frequent school information tasks on the web directory Yahooligans![1]. This web service provides a directory structure in which users can browse from a large collection of websites. A search box is also provided to let users formulate search queries to find websites matching the query terms. Bilal and Watson [1998] found that children tend to ignore the browsing categories and that they start their search directly using the search box utility. In the search box the mechanical problems identified in the previous research on digital libraries were also observed [Broch, 2000; Nahl and Harada, 1996]. The search effectiveness was hampered by the misspelling problem of the users, the lack of understanding in the use of logical operators and the formulation of queries using natural language, which were not treated adequately by the search services studied. It was also pointed out that certain queries lead only to a small amount of appropriate content for the age of the users.

Bilal [2000, 2002] carried out a follow up study of search behavior and usage of Yahooligans! with a sample of 17 users from 11 to 13 years old. Children were asked to solve open and well-defined informational tasks under two scenarios: informational task designed by the researchers and self informational tasks in which children were allowed to freely conduct their searches. In general, children were found to have limited success with the tasks given their lack of developed search skills and mechanical problems. Children also had trouble selecting the right categories in Yahooligans! In terms of browsing behavior children rarely explored thoroughly the results returned, they were found to have a search looping behavior in which previous seen results were often accessed again, and the back button of the browser was frequently activated.

[1]Today known as Yahoo! Kids: http://kids.yahoo.com/

The authors also observed a lack of engagement when carrying out well established tasks, which also hampered their search experience. On overall, children lacked of focus and seemed disoriented during the search process. The authors also pointed out that the design of Yahooligans! is not well suited for children of the age studied. Children were found to perform better under the second scenario (self-assigned tasks), in which they showed a higher tendency to engage a navigational approach, instead of a keyword search approach. The authors argued that this occurred due to the poor keyword search capabilities of the system and the greater engagement level of children when they define their own search goals.

Bowler et al. [2001] studied the search process of a small group of children aged 11 to 12 to solve school informational tasks on the Internet. The authors reported that the search engines employed (Excite, AltaVista and Yahoo!) contained information for all audiences, which discourage the users since they took long time periods to find useful pieces of information. The authors argued that the overwhelming volume of information delivered for each query led users to dead ends and the visit of previously accessed link. Additionally, children were observed to trust blindly in the results returned by the search engine making more difficult for children to assess the quality of the results.

Druin et al. [2009, 2010] characterized the search roles that children (aged 7 to 11 years old) adopt during the search process and studied how these roles depend on the children's environment and their motivation. They found that the computer expertise and orientation to explore visual content varies not only between children but also within the type of the information task. Kammerer and Bohnacker [2012] found similar trends in a recent study with 21 children aged 8 to 10. In their study, children were asked to engage informational tasks in the Google search engine. They found that children only used few keywords, which often led to an ambiguous set of web results mixing content that is suitable and non-suitable for children. They also observed that the search performance improved when using queries that are more specific.

Fidel et al. [1999] conducted a case study with eight teenage users (aged 16 to 18) in a high school library. Users were giving school assignments to be solved with a search engine. No special training was provided for the task. No restrictions were established for the search engine, and users were allowed to use their favorite search tool. Most the students opted blindly for the search engine automatically adopted by the Internet browser. The users in this age range also ignored the category browsing functionality of some search engines and favored the submission of queries. Nonetheless, these users were found to perform poorly when searching for information and they were found to reuse keywords, to have poor spelling in the formulation of queries and to revisit previous website,s even if they were off-topic for the search task. In the same line, Gunn and Hepburn [2003] observed the search information

strategies and general usage of search engines of twelfth grade students (users aged 17 to 18 years old). They found a mismatch between the self-perception of search skills by these users and their actual performance in finding quality information. The users reported themselves as good web searchers, however they were unaware of the usage of boolean operators and other mechanism to refine the search. Surprisingly most users were also unaware of search engines mechanism to narrow the search to other media type such as images.

Jochmann-Mannak et al. [2010]; Jochmann-Mannak et al. [2012] evaluated the preferences of children towards web pages with several layouts designed for children. The authors also compared search engines designed for children against Google. The case study was carried out with a group of 32 children between 8 to 12 years old. Surprisingly, they found that children tended to prefer the Google like interface to carry out their searches on the Internet. The authors found that browsing interfaces designed around children metaphors were not of the like of these users and that in rare cases these interfaces added value to a Google like service.

On overall, most of the previous studies discuss the search problems caused by the mechanical and cognitive skills of children when searching on the Internet and the mismatched between current search engines and children search capabilities. Even systems and websites that are aimed at children were not satisfying when assigning specific search tasks, as it was the case with Yahooligans! [Bilal, 2000, 2002]. Opposite results have been shown in terms of the search approached preferred by children between 8 to 12 years old. In Yahooligans! [Bilal, 2002] it was found that browsing search was preferred over keyword search, however in most of the other studies the opposite was reported [Jochmann-Mannak et al., 2010; Jochmann-Mannak et al., 2012]. From the results of these studies, it seems that even though children tend to prefer the keyword search environment, they performed better on the browsing style search given that these systems mitigate the mechanical skills of children towards spelling and query formulation.

It is important to mention that all the studies mentioned so far consider only a small group of users and focus on a specific age range. Our work differs from theirs in that we quantify the search characteristics and search difficulty of children based on aggregated results of thousands of users across a broad age range, unobtrusively, which makes our observations more representative on a web-scale. Additionally we report topic interest trends over a population with diverse demographic characteristics, which is not possible to observed with a limited number of users. Moreover, no study mentioned so far contrasted the search behavior between age ranges, which we address in this chapter using fine-grained age ranges of young users. Another important research concern that is not addressed in any of the studies mentioned is the understanding of the activities that children carry out on the Internet browser outside the scope of a search engine, and how these activities motivate search in

young users. We address this research gap through the analysis of the Yahoo! toolbar logs in Chapter 4.

2.2.2 Related query log analysis

Several studies have been carried out to analysis large-scale query logs of commercial search engines. Silverstein et al. [1999] analyzed a query log of the Altavista search engine that contains approximately 1 billion entries. They presented an analysis of individual queries, query duplication, query sessions and correlations between query terms based on a set of descriptive measures such as query length, query frequency, session length and term frequency. According to this study, users tend to utilize short queries (mean of 2.3 words per query) and user sessions are short on average (2 queries per session). They also reported that queries are not changed often by users and that 77.5% of the queries are unique, which suggests a wide variety of information needs and several ways to express them. Similar results regarding query length and query characteristics are reported in the analysis done by Spink et al. [2001] on a smaller query log of the Excite search engine.

Pass et al. [2006] analyzed various aspects of an AOL query log such as query formulation patterns, search engine efficiency, user demographics and user's interactions. They described the query space as vast, topically diverse and in constant change. Interestingly, they also found that 20% of the users perform approximately 70% of the queries and that less than 1% of the web domains account for more than 50% of the clicks of the users. Further analysis on this query log is carried out by Brenes and Gayo-Avello [2009] by grouping the queries and sessions based on the query popularity. They found different behaviors throughout the groups (i.e. navigational coefficient, query length, temporal length) which suggested that more fine-grained analysis are required to study query logs.

A crucial aspect to study query logs is the definition of the user session. A session is a sequence of queries issued by the user to satisfy an information need. Boldi et al. [2008]; He and Goker [2000]; Jones and Klinkner [2008] construct search sessions using a time-out cut-off between queries, which establishes that two queries are in the same session if the time difference in which they were issued is smaller than a given threshold value. We will employ an analogous definition of search session in this and the following chapters.

2.3 Research Method

Search log analysis is one of the main research methods in web search studies to capture unobtrusively the interaction of a large number of users with a search engine [Rieh and Xie, 2006]. These analyses have been used extensively to generate statistics

of search engine usage, evaluate web site designs, and test theoretical hypotheses about the effects of different search engine functionality based on user behavioral data [Burton and Walther, 2001; Rieh and Xie, 2006].

We employed the method *TLA* (transaction log analysis) to conduct web search transaction log analysis [Jansen, 2006]. This method was designed for studies aiming at understanding the interactions between searchers, the system and the content provided by these systems. The objective reached through the understanding of the interaction of these three factors may involve improving the design of a search system, improve its accessibility or identifying the user's searching behavior [Jansen, 2006]. The interactions refer, within the scope of this method, to a "mechanical expression of underlying information needs" [Jansen, 2006] and correspond to the communication between the searcher and the system.

TLA is based on a grounded theory approach, in which a systematic discovery of theory is carried out through sampling, comparing and analysis of data. The resulting models are derived in an inductive manner since the results are grounded in observations of the *real world*, instead of being generated by abstract constructs [Jansen, 2006]. In TLA is explored the characteristics of the search sessions and trends are identified from the queries submitted by the user, how the queries are modified within the search sessions, how the result list is explored, and based on the type of content that is explored (e.g. multimedia, plain text).

This method consists of three steps:

1. Collection: Registering and gathering log data within a predefined time window.

2. Preparation: The data cleaning process carried out to reduce noise in the data and disregard irrelevant entries for the study.

3. Analysis: The process of extracting a set of metrics from the data prepared, and analyzing the metrics in respect to a set of research questions.

In the following section is described the collection and preparation step.

2.4 Data Collection and Preparation

The AOL query log contains approximately 36 million entries and it was collected during two months in 2006. Currently, this is the largest and most up-to-date freely accessible query log on the web. Each entry contains an anonymous user ID, time of submission, rank position and domain of the URL clicked. It is important to mention that there is controversy about the usage of this query log in the research community given the privacy issues that arose by the identification of actual users in the press [M. Barbaro]. Nonetheless, all the results presented in this thesis are obtained from

accumulated counts and no actual user identification is performed. The identification of the queries employed to retrieve content for children was performed by matching the entries listed in the *Dmoz* kids & teens directory, which contained 45.635 entries, with the domains clicked in the entries of the query log. Given that the query log does not include the entire URL visited, matches are restricted to the cases in which only the domain is listed as a *Dmoz* entry. Three data sets of query log entries were constructed by employing the matching procedure described above on the *Dmoz* entries tagged for kids, teens and mature teens.

Sessions are constructed by grouping contiguous queries submitted with a time difference smaller than t_θ and that are from the same user. A formal definition of session is shown in Equation 2.1.

$$S = \langle \langle q_{i_1}, u_{i_1}, t_{i_1} \rangle, ..., \langle q_{i_k}, u_{i_k}, t_{i_k} \rangle \rangle \tag{2.1}$$

where $u_{i_1} = ... = u_{i_k}$, $t_{i_1} \leq ... \leq t_{i_k}$ and $t_{i_{j+1}} - t_{i_j} \leq t_\theta$ for all $j = 1, 2..., k-1$ The parameter t_θ was set to 30 minutes because it is the most common value employed in the literature [Boldi et al., 2008; He and Goker, 2000; Jones and Klinkner, 2008]. We consider that this time window is also suitable for sessions expressing children information needs, since it has been shown that the average time that children spend to fulfill an information need varies between 10 and 16 minutes [Bilal, 2002]. The sessions used to satisfy children's information needs are those that visit at least one *Dmoz* domain.

In terms of data cleaning we removed those log entries satisfying any of the following criteria:

- Queries consisting only of the character "-" or with a length greater than 20 tokens.
- Queries consisting of only non-alphanumerical characters.
- Log entries in which the click occurs beyond the 40th ranked position.
- Search sessions lasting more than 24 hours.

The restriction were applied to remove spam and abnormal search activity. In regard to the first restriction, we observed that several queries consist only of the hyphen character ("-"). This may be due to a preprocessing (perhaps anonymization process) applied to the AOL log before it was released. We simply discarded these entries.

Three sets were built using the search session definition explained above and the cleaning criteria steps. Recall that each data set is constructed based on the age tag of the urls: *kids* (users up to 10 years old), *teens* (users aged 13-15 years old) and

	Queries	Uniq. Q.	Sessions	Goals
Kids	485.861	10.252	21.009	32.292
Teens	411.474	4.169	7.930	14.503
M.Teens	516.570	10.057	15.519	26.600
All log	36.389.577	10.154.747	10.769.830	8.005.597

Table 2.1: Size of the query sets.

mature teens (users aged 16 to 18 years old). The group of *mature teens* will be referred as *m.teens* in the results reported. Table 2.1 summarizes the characteristics of the sets collected.

2.5 Analysis

All the statistics report in this chapter are micro-averages. For instance, for the case of query length we count the number of tokens in each query and then we average within each query set (*kids*, *teens*, *mteens* and *all* set). Alternatively, these metrics can be estimated using macro-averages. In this case, the average query length for each user is estimated and then these values are averaged across all users to create an overall estimate for each set.

In this chapter we opted to report micro-averages. We believe this approach is more sound for the analysis of this chapter given that the log entries were chosen based on a set of urls and not based on demographics of the users from the search logs.

In the following chapters we will employ macro-averages instead of micro-averages because we have a greater certainty that the users are actual children or teenagers, moreover the data set is collected based on the age of the user (which is known for the search logs used in the next chapter). Nonetheless, to ease the comparison between the results obtained in this chapter and the next chapter, we also report in Appendix A macro-averages for most of the results presented in this chapter.

In general terms, although we found slight differences for each metric between the micro and macro-averaged values, we did not find differences in the trends observed across the four age sets and consequently the conclusions and observations drawn from the micro-averaged results also hold on the macro-averaged results. These results are reported in Appendix A for completeness.

Statistical Test

In this chapter we address *RQ-1.1* and *RQ-1.2* by comparing averages across the four data sets for each one of the metrics considered. The differences between the averages are tested for statistical significant using the two-tailed t-test for the equality of means

with unequal variance and sample size[1]. We consider a difference to be statistically significant if the probability of the null hypothesis, i.e. the two means being equal, is smaller than 0.1%. It is important to clarify that due to the large volume of data involved in the analysis, most differences were found statistically significant at a 0.1% level according to the two-tailed t-test. The p-values observed for most of the test were smaller than 0.001 which, lead to the rejection of the null hypothesis of the statistic under consideration. Nonetheless, we will state in each section the cases in which a given statistic (e.g. average query length) was not statistical significant when comparing groups.

2.6 Query level results

In this section is explored the characteristics of the query entries. We considered the query length (which is measured by the number of words per query), the type of queries (question queries, phrasal queries, query intent), cue words, the rank position of the domains clicked, the query frequency distribution and the distribution of users across the datasets.

2.6.1 Query Length Analysis

The formulation of a well-defined query is a crucial part of the search process in IR systems [Downey et al., 2008]. The correlation between query length and IR effectiveness has widely been explored before [Belkin et al., 2003; Jansen et al., 2000]. On TREC ad-hoc settings, it has been found that long queries (i.e. longer than average) lead to better search performance and user satisfaction [Belkin et al., 2003]. Nonetheless, recent studies show that this result does not always hold on the web scale [Downey et al., 2008].

Recently, a strong association between the length of the query and the specificity of the user's query intent [Phan et al., 2007; Roul and Sahay, 2012] has been found, in which longer queries lead to a more specific and less ambiguous set of results. Thus, the submission of longer queries to the search engine is a strong indicator of the capacity of the user to construct queries that are more specific.

The average query length found for the kids, teens and mature teens datasets were 3.8, 3.4 and 3.2, respectively. These values differ from the mean of the entire query log, which is 2.5 words per query. Figure 2.1 depicts the query length distribution for each set. All pair comparisons were found statistical significant according the test describe in Section 2.5. The average query length of the entire data is also in line with the average length reported for other large scale query logs [Bendersky and Croft, 2009; Silverstein et al., 1999].

[1] `http://en.wikipedia.org/wiki/Student_t_test`

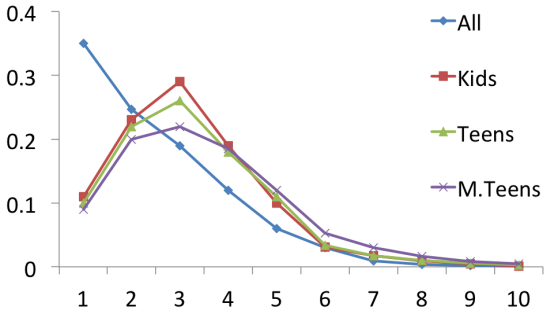


Figure 2.1: Query length distribution.

This result may look surprising at first given that young users are expected to have more difficulties formulating specific queries. Similarly young users have a smaller vocabulary which is reflected in the vocabulary used in the queries, as it will be shown in Section 2.6.5. However, this result may be due to the fact that young users have been observed to user more natural language in their queries. Additionally, these users can have more aid from their parents, which explains why the queries for the kids and teens data sets are more elaborated (i.e. longer). In the next section we will explore the usage of greater natural language in these queries and some cues suggesting greater parental supervision.

Given that the queries analyzed were chosen through a set of urls, longer queries in these type of queries also shows that focused queries are required to access high quality content oriented for children and teenagers.

2.6.2 Natural Language Usage

In previous studies [Druin et al., 2009] children aged 8 to 12 years old have been observed to have a tendency to write queries using natural language constructions more often than adults, especially when the information need requires multiple phases to be solved. It has also been found that children tend to express complex information needs by directly typing the question they have [Bilal, 2002].

The following query types were created to quantify the usage of natural language in the queries.

1. Question queries: Queries for which the first token is a question word (e.g. how, where, what) or the last character of the query is a question mark (e.g. what is the only immortal animal?)

2. Modal queries: Queries containing auxiliary verbs as will, won't, don't or modal verbs as shall, should, can, etc. (e.g. I don't want to go school)

	question	modals	know. quest.	superlatives	kids focus	total N.L.
Kids	3.9%	4.8%	0.6%	0.3%	2.4%	9.6%
Teens	3.7%	2.7%	0.4%	0.3%	0.1%	7.1%
M.Teens	3.9%	4.1%	0.2%	0.3%	0.1%	8.6%
All	2.7%	1.7%	0.1%	0.3%	0.0%	4.8%

Table 2.2: Natural language usage. Frequencies of the natural language construct defined are displayed for each data set. The kids focus group refers to the *kids targeted queries*. *Total N.L* refers to the sum of all the types except for the *kids focus* type.

3. Knowledge questions: Queries containing the words *describe*, *about*, *explain*, *define* or *interesting*

4. Superlatives: Queries containing superlative adjectives (e.g. the fastest dog)

5. Kids targeted queries: Queries containing any of the following expressions: *for kids*, *for children*, *4 kids* or *4 children*.

Knowledge queries attempt to measure the fraction of queries intended to fetch a specific explanation about an issue or topic. We decided to explore the usage of superlative queries because this types of construct is commonly employed by children to satisfy their curiosity about certain topics, such as in the query"fastest animal". Note that queries with superlative constructs target in a more concise way a particular aspect of the surrounding world (i.e. objects, situations, persons) compare to other natural language constructs such as adjectives or comparative adjectives [Cecchin, 1987]. Superlative queries were detected by looking at tokens with the suffix *est*, and filtering out those matched tokens that are not listed as adjectives in Wordnet[1], or that have a locational meaning (e.g. west).

The kids targeted queries were included to identify if children of different ages were using this mechanism to focus the query on content that is suitable for them. The frequency of these type of queries are shown in Table 2.2. In this table is shown that question queries are more frequent in the children set (queries used to retrieve information for children) than in the general purpose query set, which is in line with Druin's observations [Druin et al., 2009]. However, we did not find statistical difference between the kids and mature teens query sets.

No particular trends were observed for the superlative type of queries (no statistical differences were found). We observed a slightly higher usage of knowledge queries in the kids and teenagers data sets. However, no statistical significance were found when comparing the teens, mature teens and *all* query sets. A greater usage of modal constructs was also found in the kids and teens data sets. Particularly the kids queries were observed to have the greatest percentage of this type of queries (4.8%

[1] `http://wordnet.princeton.edu/`

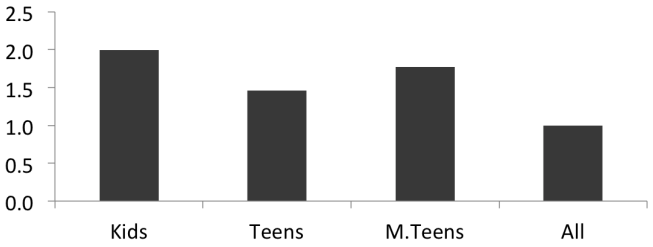


Figure 2.2: Ratio of natural language usage in queries for each set against the *all* query set. This figure shows that users from the *kids* set submit twice as much queries with natural language constructs than users from the *all* set.

against 1.7% for all type of queries). The teen queries also have a higher percentage of this type of queries in respect to adults (2.7%).

We did observe a higher usage of the *focus* query type in the kids data set. This result indicates that these users were aware that the content targeted was for a special and specific public segment on the Internet. However, this result may just suggest parental supervision, since adding the suffix *for kids* to the query is a reasonable strategy for parents to follow when searching for information for their children.

On overall, we observed a greater usage of natural language in the kids, teens and mature teens datasets. Figure 2.2 shows the ratio between the percentage of queries with natural language for each set in respect to the adults queries (i.e. *all* query set). The percentage of natural language is estimated by summing up all but the *focus* query type. The kids queries contain up to twice as much natural language in respect to all the queries in the data set. For the case of teenagers the ratio was around 1.5.

The differences found in this section are summarized as follows:

- Significant larger number of question queries in the *kids* and both *teenagers* (*teens* and *mteens*) sets in respect to the average web user.
- Only marginal differences between the *kids* and *teenagers* query sets in the proportion of question queries
- Focused queries clearly preferred in the *kids* query set.
- No clear trends found for the superlative query types.
- Modal queries preferred in the *kids* and *mature teens* query sets.

2.6.3 Query intent Analysis

Queries were also analyzed using the classification suggested by Broder [2002], which captures three types of user intent: informational, navigational and transactional

	All	Kids	Teens	M.Teens
Informational	51.7%	33.5%	51.05%	49%
Navigational	21.5%	13.0%	32.7%	33.2%
Transactional	26.7%	53.3%	16.2%	17.5%

Table 2.3: Query intent proportions.

queries. Informational queries are used to address an information need by locating content relevant to the topic of interest (e.g. areas in africa giraffes live in). Navigational queries are used to locate a specific website, which can be the main Website of an organization or a hub site (i.e. bobthebuilder.com). Transactional queries are used to locate a Website with the aim of obtaining a product. The product may refer to an item to be purchased, an application to be executed on-line (i.e. alphabet coloring pages) or a multimedia resource to be downloaded.

The results summarized in Table 2.3 were obtained by manually classifying the queries using the guidelines given by Jansen et al. Jansen et al. [2008] to classify query intent (Broder's categories are also used in this study). The queries were classified by sampling randomly at 15% the unique queries of the kids, teens and mature teens data set. For the whole data set, a random sample of 1400 queries was obtained. This size is comparable to the size of the sample employed in a previous study on a large query log [Broder, 2002].

We found that informational queries are preferred in the teens and mature teens and *all* set over transactional and navigational queries. Previous studies have also found this behavior on large query logs. For instance Broder [2002] reported on a random sample of 1000 queries from the Altavista log that 48% of the queries were informational, 30% transactional and 20% navigational, which are comparable to the percentages obtained in our sample.

Interestingly, this trend was not observed for the kids queries, in which transactional queries are preferred (increase of 20% in respect to the average web user). We found that transactional queries are mainly used in the kids and teens queries to interact with web applications (e.g. flash/java games, academic quizzes) or to obtain free on-line resources (e.g. poems, songs lyrics, coloring pages).

Conclusions in the same line have recently been drawn in Ofcom [2010]. They reported that gaming is the preferred Internet activity for children aged 5-7 and second preferred for children aged 8-11. In particular 37% of the children aged 5-7 (52% for children aged 8-11) were found to use the Internet at least once per week for gaming, while 19% use it for information purposes (46% for children aged 8-11). On the other hand, for the case of users aged 12 -15, informational and social activities are more popular than gaming. In this case 66% use the Internet at least once per week for informational purposes and 48% for gaming.

The lower use of informational search in the kids queries compared to the other

query sets can be caused by the current lack of specialized IR applications to satisfy children's information needs or to the unsuitability of most of the content on the web for these users. Given that these users are more familiar with the interaction of multimedia and on-line applications, the design of more interactive tools can highly improved the motivation and success of users searching for children-friendly information.

2.6.4 Cue words analysis

Cue words provide information about the characteristics of the content searched by users and their identification have been proved useful to aid search through query expansion techniques [Guo and Ramakrishnan, 2009]. The motivation of this section is to identify the most common characteristics and content type of the information searched in the query sets. The identification of cue words is based on a contextual model [Wang and Zhai, 2008], which is defined in Equation 2.2.

$$P(a|w) = \frac{cf(a, C(w))}{\sum_i cf(i, C(w))} \tag{2.2}$$

where $C(w)$ is the context of w . This equation models the likelihood of a word a to appear in the context of a given word w. In this thesis, we define the context of w_i as the set of words that occur at positions $i-2, i-1, i+1, i+2$. We are interested in mining the context words (cue words) used in the query sets. For this purpose, the words in the queries are clustered such that each cluster represents the set of content words w that co-occur with a given context word. Thus, each cluster can be seen as a group of information needs that make use of the same cue word.

The star algorithm is used to perform the clustering by representing the nucleus of each cluster as the context word [Gil-García and Pons-Porrata, 2008], the satellites as the content words co-occurring with the context word, and the similarity measure between two vertices as the probability given by the model defined in Equation 2.2. Table 2.4 shows the top 5 clusters ranked by size for the three query sets. We found 56, 67 and 69 clusters in the kids, teen and mature teens queries, respectively. Only 5% percent of the cue words of the kids query set appear as cue words in the teens query set, and only 8% for the case of the teens and mature teens query sets. We observed a clear relation between the clusters and the expected topics of interests of the target users and progression in the topics. For instance social related topics (e.g. life, myspace, boyfriend) are found in big clusters in the teens and mature teens but they are not present in the children's clusters. The differences observed are summarized in the following manner:

	Cue word	Content words
kids	coloring	pages,free, page, easter, book, day, bible, disney, pring, preschool, butterfly, mother's, animal...
	craft	ideas, tissue, projects, kite, luau, invitation, frame moms, folded, gallon,plastic, shells, canoe...
	poems	funny, teachers, short, haiku, silverstein, concrete, shel, seasons, appreciation, mother, acrostic...
	help	thank, neopets, abc's, fluoride, teeth, insects, text, myspace, ways, magnesium, hw, guild...
	game	mystery, files, dressup, match, pool, board, sites, play, create, yahtzee, maze, math...
teens	ps2	cheats, games, game, godfather, codes, 2006, andreas, baseball, mlb, 2k6, naruto...
	videos	drift, jupiter, randy, snakes, orton, stunt, hero, racer, angel, d1...
	facts	fun, pearl, harbor,abstinence, planets, aids, jazz, yeast, thailand...
	school	statesmen, greenbriar, rule, freshman, houses, centerpieces, jones...
	life	wonderful, important, called, teenagers, george, factor, survival, character, roaring...
m. teens	body	muscles, odor, stay, infections, flexible, water, piercings, development, bruises, collagen...
	news	fox, science, channel, nasa, bahrain,latest,soft, knoxville, superstring, singel, dinosaur, iraq...
	download	orion, beholder, tv, shows, boggle, dominoes, fmv, craps, snood, knight, scrabble, collapse...
	map	middle, east, israel, egypt, okinawa, galilee, sea, jerusalem, palestine, detailed, surrounding...
	college	grants, scholarships, sinclair, community, search, darton, kissing, binge, drinking, financial...

Table 2.4: Cue clusters.

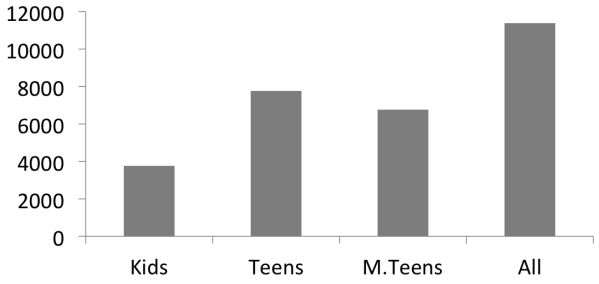


Figure 2.3: Query vocabulary of each data set.

- Different content genres are prominent in each data set. For instance *coloring pages*, *videos* and *news* in the *kids*, *teens* and *mature teens* sets respectively.

- The game cluster was one of the largest for both the *kids* and *teens* data sets. However, cue words in both clusters exposed different aspects associated to these clusters. In the former online and educational games are searched while in the latter commercial/console games are preferred.

- Similarly, the school cluster was found for all the query sets. For the *kids* queries the focus of the keywords composing the cluster is on online help for school tasks. For the teen queries the focus is on factual information, while for the *mature teens* queries the focus is on preparation for the college years.

- Social information aspects are prominent for the *teens* and *mature teens* data set.

2.6.5 Query vocabulary size analysis

The vocabulary size has been employed in the field of language acquisition as a predictor for reading comprehension and communication [Hellman, 2011]. This metric has also been employed to track language development from early childhood to adulthood [Tonzar et al., 2009]. From the engineering point of view, the vocabulary size has a deep impact in the design of systems for machine translations, speech recognition and part of speech taggers [Church, 2011]. In terms of web search, we believe that the vocabulary size can be interpreted as an estimator, in a broad sense, of the capability of the users to understand the content of websites, and in the case of web queries, as the capability to retrieve content from the Internet through web queries. Thus, a bigger vocabulary size indicates a greater capability to find relevant information. This interpretation is analogous to the capability of carrying out communication, which is associated to the size of the vocabulary in the case of natural language.

We estimated the vocabulary size by counting the number of distinct words employed in a sample of *10K* queries. To make the comparison fair across age ranges we employed a uniform random sample of the same size for each one of the age groups studied.

Figure 2.3 shows the vocabulary size results. We observed a significantly smaller size for the case of kids queries (3.7K) in respect to the teenagers (7.7K) and for the sample obtained from the whole data set (11K). For the sample obtained from the mature teen queries, we obtained a slightly lower vocabulary size than the one obtained for the teens query set (6.7K). This may be due to the focus of the Dmoz Kids and Teens directory on younger users. This result suggests that query expansion functionality are highly beneficial for the case of queries targeting content for children since the vocabulary is significantly smaller. This result also shows that these queries are able to target a significantly smaller portion of the Web.

In the next cshapter we will contrast these results against the vocabulary obtained from queries submitted by young users.

2.6.6 Topic Distribution analysis

A topic distribution was associated to each query set using the category hierarchy of the Yahoo! Directory[1] and a sample of 11 thousand queries from each set. Each query in the sample was mapped to one of the categories in the Yahoo! Directory. Previous to the sampling process, we removed navigational queries in the sets using straight forward rules to detect this type of queries (i.e. removing queries containing urls or domain names).

We carried out this analysis using the main categories of the directory and one level of depth (i.e. subcategories) in the hierarchy. The mapping is carried out by performing a Google site search using the Yahoo! Directory url and then applying majority vote to select the category associated to the query.

Figure 2.4 shows the micro-averaged distribution of topics for each one of the datasets. It is important to mention that we were not able to classify the 19%, 20% and 14% of the queries in the sample of kids (and teens), mature teens and whole query set. The proportions shown in Figure 2.4 were estimated by normalizing all the topics retrieved to 1 (after discarding the non-classified queries). Only the most representative topics in the distribution are shown to improve the figure readability. An analogous macro-averaged plot can be found in Appendix A (Figure A.5). In general terms the results in both figures are very similar and all the trends are preserved.

Expected trends were observed in Figure 2.4. For instance a clear decreasing trend of queries aiming for *recreational/games* were found. The 29% of the kids queries target this type of content against 2.8% for the queries sampled from the entire query

[1] http://dir.yahoo.com/

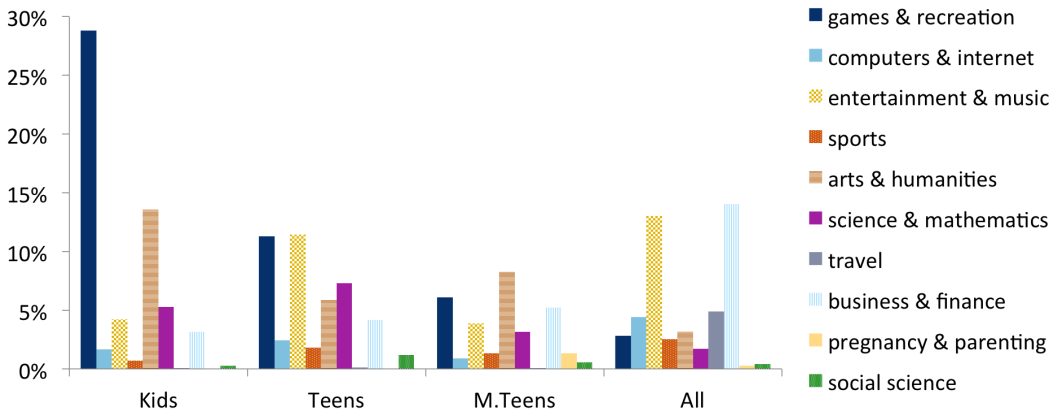


Figure 2.4: Topic distribution of each data set (Yahoo! Directory categories).

set. The *art* topic also has the largest proportion for the case of the kids queries and it decreases through the other query samples. For the case of teenagers a large fraction of *entertainment*, *science* and *games* were observed. In regard to the sample derived from the whole data set, we observed that topics such as *finances* and *travel* gain importance. It is important to mention that the distribution found for the entire data set is more disperse than the distribution found for the kids and teens query sets, that is, a larger number of topics was observed and lower percentages for the most prominent topics. For the kids and teen queries was found high value proportions focused on a smaller number of topics.

On overall, these results show that the queries extracted through the Dmoz urls are a fair representation of the topic interests of the users targeted (i.e. young users).

In the next chapter we will come back to this analysis. We will measure the correlation between the distributions obtained in this chapter and the distribution of topics associated to the queries submitted by young users using search logs from the Yahoo! search engine.

2.6.7 Click analysis

The click information of the datasets was analyzed to compare the retrieval performance between children, teenagers and general purpose content queries. For this analysis we collected the rank distribution, the mean reciprocal rank (MRR) and the click frequency of the queries in the datasets. Queries leading to clicks on high ranked positions indicates that the information needs can be satisfied efficiently by the IR system. The MRR of a set of queries Q is defined by Equation 2.3. Lower MRR values occur when lower ranked documents are frequently clicked by the user, which

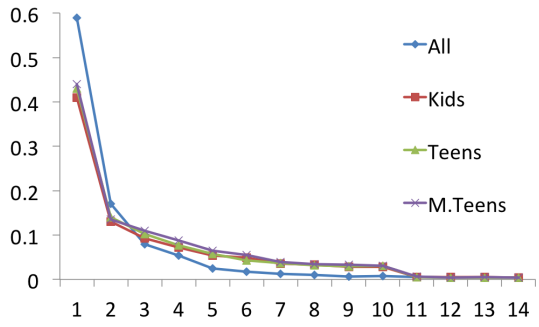


Figure 2.5: Rank Distribution.

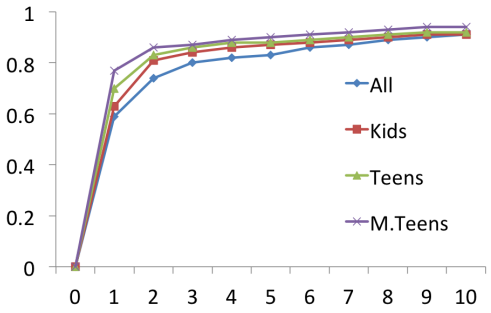


Figure 2.6: Query frequency distribution.

indicates poorer retrieval performance.

$$MRR = \frac{1}{\mid Q \mid} \sum_{i=1}^{|Q|} \frac{1}{rank(Q_i)} \tag{2.3}$$

Figure 2.5 shows the rank frequency distribution of the data sets. This figure demonstrates that the retrieval performance of the queries to retrieve children information is poorer than the queries used to retrieve non-children oriented content, since clicks on lower ranked results are more frequent in the kids and teens query set. The MRR values found for the kids, teens, mature teens and whole data set were 0.57, 0.59, 0.58 and 0.73, respectively, which lead to the same conclusions since the MRR value for the whole data set is greater than in the other sets.

This result also shows that high quality content designed for children and teenagers are less accessible than general purpose content. Particularly, it also shows that a bigger effort is required from the user to access this type of information, since they have to click on lower ranked results. Note that this result is consistent with the longer average query length found in the kids and teenagers data sets since the usage of longer queries suggests the need of more elaborated queries to find content suitable for this niche of users. All results were found statistical significant between the four data sets.

2.6.8 Query frequency analysis

Figure 2.6 shows the cumulative frequency of the queries in our data. We observed a similar behavior in all the data sets since queries that appear up to three times in the query log account for 80% of the total number of queries. However, a greater number of unique queries with low frequency are found in the children and teenager query sets. This may be due to the fact that users finding information for young users have more problems finding appropriate content, which causes the formulations of more

Kids	Teens	M_Teens
nickjr.com	the n.com	prom hairstyles
elmo	nasa	american idol
nick jr	hairstyles	cheats
coloring pages	kingdom hearts 2	cheat codes
postopia	claires	prom dresses
candystand	celebrity hairstyles	pussycat dolls
the wiggles	christina aguilera	ea sports
starfall.com	degrassi	bladder infection
dora the explorer	gurl.com	scholarships
primary games	homestarrunner	game cheats

Table 2.5: Most frequent queries.

	Kids	Teens	M_Teens
Kids	100.0%	6.0%	9.2%
Teens	13.8%	100.0%	20.3%
M_Teens	10.7%	10.3%	100.0%

Table 2.6: Proportion of users across the datasets.

queries as attempts to fulfil the information needs.

Table 2.5 shows the 10 most frequent queries in the data[1]. As it is expected, these queries reflected typical interests of the user target groups. For instance, *elmo* and *dora the explorer* are very popular characters among children. The teens and mature teens queries show a greater interests in video games (*ea sports*), social events (*prom*) and educational matters (*scholarships*), which indicates differences in the information needs between children and older users.

2.6.9 User analysis

We analyzed the proportion of users that submit queries in more than one of the data sets. The results are summarized in Table 2.6. This table shows that users that retrieve information for children rarely submit queries to extract information from the teens and mature teens data sets. Analogous results were found for the other query sets.

2.7 Session level results

Sessions allow us to understand the way users accomplish information needs and how they interact with the search engines. We collected three types of metrics to

[1]Variations of the same domain were removed from this list (e.g nickjr and nickjr.com).

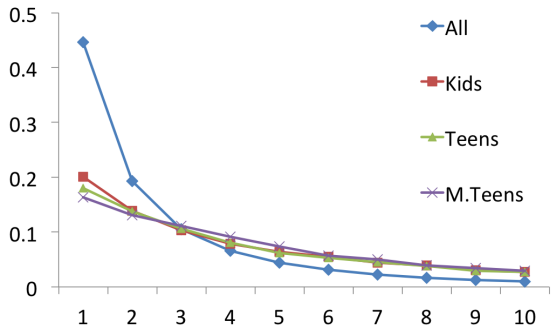


Figure 2.7: Sessions length distribution.

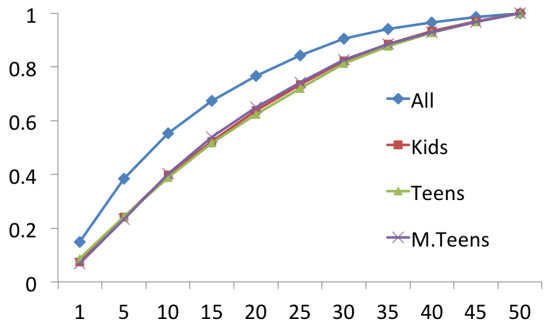


Figure 2.8: Session duration distribution.

	All	Kids	Teens	M_Teens
S. length (mean)	3.3	8.7	10.1	9.8
S. length (median)	2.0	5.0	5.0	6.0
S. duration (mean)	11.7	20.4	23.1	21.5
S. duration (median)	5.1	12.5	14.6	12.7

Table 2.7: Summary of sessions characteristics.

compare the session characteristics of our datasets: session length, duration and query reformulation metrics. Session length is the number of entries issued in the session. Entries can refer to new queries issued by the user or to results clicked using the same query. This metric is an indicator of search efficiency since a greater amount of entries suggests that more changes to the queries and document visits are needed to fulfill the search task. Session duration is defined as the time in minutes between the last and first query issued in the session. This metric is an indicator of the complexity of the underlying information need. Query reformulation metrics can be used to understand the way users change their queries to reach their information needs. Table 2.7 summarizes the metrics obtained for our data sets. All the results obtained were found statistically significant.

2.7.1 Sessions length

Figure 2.7 shows that sessions from the average web user are mostly short given that 80% of the sessions contain less than 5 query entries. On the other hand, the length of the sessions used to retrieve information for young users tend to be longer and its distribution is more dispersed. The average length found for the average web user sessions are in line with previous studies, in which the estimated average session length was 2.8 [Jansen et al., 2008; Silverstein et al., 1999]. The longer average length found for the other sessions (see Table 2.7) suggests that these users weren't certain of the relevance of the information found, since they had to submit more queries and

explored more results.

This result can also indicate that the documents retrieved by the search engine are not sufficient to satisfy the user's information needs. This result is consistent with previous findings [Bilal, 2002], in which children showed *non-linear navigation style* when solving research tasks. This search style is characterized by the exploration of several choices before a final relevance judgment is made [Bilal, 2002].

2.7.2 Sessions duration

The duration in minutes of the four types of sessions is shown in Figure 2.8. This figure shows that users required more time to explore and complete information needs associated to content for young users, than to content for general-purpose content. Statistically significant differences were found between the kids/teenagers/mature teenagers and general-purpose sessions. No significant differences were found among the types of young user sessions. The longer duration of sessions to retrieve information for young users suggests more difficulty to solve the information tasks associated to these sessions and the greater difficulty to access high quality content for this type of users. This result is consistent with the greater amount of pages visited and queries issued found in the session length analysis. The trend of clicking on higher ranked pages observed in the previous section for the kids and teens data sets is also in-line with the results show in this section.

We estimated the session success rate by considering as *successful searches* those sessions in which the last event is a click, as it is suggested by Brenes and Gayo-Avello [2009]. We found that the success rate for the kids, teens, mature teens and whole data set were 82.5%, 79%, 79.8%, 52.2%, respectively. These values would indicate that users of the young query sessions are more successful in their information seeking tasks. However, we consider that this metric can also be indicator of the *trust* that young users have on web results, in which more clicks indicate more *trust*. In [Ofcom, 2010] is reported that only 49% of the children (aged 12-15) make some critical judgment about the truthfulness of the results.

2.7.3 Query reformulation analysis

Users constantly modify their queries in an attempt to get better results from the search engine. The analysis of these query refinements allow us to have a better understanding of the way user's interact with the search engine and the search strategies employed to satisfy their information needs. In this chapter we analyze the following types of query reformulations:

- Words added to the query (w.a): The previous query is a strict suffix of the target query. For instance, $\{\text{dora}\}_{i-1} \underset{w.a}{\rightarrow} \{\text{dora the explorer}\}_i$

	All	Kids	Teens	M_Teens
n.q.	48.1%	25.9%	26.5%	24.5%
w.a.	1.5%	1.4%	1.8%	1.8%
w.r.	0.1%	0.2%	0.2%	0.2%
w.c.	6.3%	7.8%	8.0%	8.5%
m.r.	38.8%	59.6%	57.6%	59.0%
p.q.	3.0%	3.3%	3.7%	3.5%
s.c.	1.8%	1.5%	2.1%	2.1%

Table 2.8: Frequency of query reformulations.

- Words removed from the query (w.r): Target query is a strict suffix of the previous query. For instance, $\{\text{barbie coloring pages}\}_{i-1} \underset{w.r}{\rightarrow} \{\text{barbie}\}_i$
- Change of words in the query (w.c): Target query contains at least one word in common with the previous query. Word order is ignored. For instance, $\{\text{all about spiders}\}_{i-1} \underset{w.c}{\rightarrow} \{\text{all about cobras}\}_i$
- Spelling correction (s.c): The Levenshtein distance between the target and previous query is ≤ 2. For instance, $\{\text{candysand}\}_{i-1} \underset{s.c}{\rightarrow} \{\text{candystand}\}_i$
- New query (n.q): Target query does not share keywords with the previous query and the Levenshtein distance is greater than 2. For instance, $\{\text{sesame street}\}_{i-1} \underset{n.q}{\rightarrow} \{\text{elmo}\}_i$
- More results from the same query (m.r): Target query is identical to the previous query and it is used to access a different website.
- Return to previous query (p.q): The target query was submitted during the same session.

Similar query reformulations types have been used in previous query log analysis [Huang and Efthimiadis, 2009; Pass et al., 2006]. Although $m.r$ is not a formal query reformulation (since no change is performed on the previous query), we included it because this action is commonly used in the search process. Table 2.8 shows the percentages of the query reformulation types found in our data. These percentages were calculated on sessions containing more than one query, which correspond to the 83%, 87%, 90% and 55% of the kids, teens, mature teens and whole data sessions, respectively. A salient difference is the average drop of 22.5% of new queries issued in the kids and teenager sessions compare to the general-purpose sessions.

Most of this drop is reflected in the greater usage of the same queries to explore further results, which accounts on average for 90% of the new query reformulation type drop. Small gains in all the other query reformulation types were also found in

the teens and mature teens sessions. Nonetheless, kids sessions only showed increase in the word removing, word changing and reusing of previous queries. Although, it has been shown that children commit spelling mistakes more frequently than older users, we didn't find an increase of the spelling correction reformulation type in the kids sessions. The rank data between query pairs and the click patterns suggested in [Huang and Efthimiadis, 2009] were employed to evaluate the effectiveness of the query reformulations. These click patterns are established between a query and its reformulation and can be described as follows:

1. A domain was accessed in the previous query and in the reformulation query (*click_click*)

2. A domain was not accessed in the previous query and in the reformulation query (*skip_click*)

3. A domain was accessed in the previous query but not in the reformulation query (*click_skip*)

4. A domain was accessed in the reformulation query but not in the previous query frequency (*skip_skip*)

Table 2.9 summarizes the click patterns found for each one of the reformulations types defined above. Patterns (1) and (4) indicate that the reformulation was successful, on the other hand, patterns (2) and (3) indicate that the query reformulation was not effective. The ratio between these patterns also provide important information about the use of the query reformulations and their effectiveness. Low values of the ratio $\frac{(1)+(3)}{(2)+(4)}$ shows that the user is not satisfied with the initial query and the query reformulation is employed as an attempt to obtain better results. On the other hand, high values of this ratio indicates that users are refining or specializing their queries since a result has already been accessed. The query reformulation types *w.c*, *w.r* and *m.r* had the highest ratio considering all the sessions of the datasets. This result shows that users tend to use these types of reformulations to refine their queries, since previous pages were accessed using similar queries.

The high ratio value found for *m.r* is logical given that users often keep the same query when previous results have been useful, as it is suggested by the high percentage found in the click-click pattern. Low rate values were found for *w.a*, *p.q* and *s.c*. Thus, these reformulation types are mostly used when users are not satisfied with their current queries. Similar conclusions were drawn by Huang and Efthimiadis [2009] on this query log by using a more extended set of query reformulations.

	Dataset	Click-Click	Click-Skip	Skip-Click	Skip-Skip	Mean rank change	Mean time change
n.q	kids	35.9%	22.1%	23.2%	18.6%	-3.8	371.2
	teens	35.7%	22.4%	22%	19.7%	-3.04	359.6
	m_teens	36.7%	22.5%	21.6%	19%	-4.11	371.6
	all	23.7%	19.64%	20.3%	36.2%	-1.98	337.3
w.a	kids	30.6%	14.8%	28.3%	26.1%	-5.83	136.19
	teens	31.6%	13.8%	27.7%	26.8%	-3.34	133.8
	m_teens	30.8%	14.2%	30.7%	24.2%	-7.29	293.8
	all	18.1%	13.2%	28.9%	39.6%	-2.9	117.4
w.r	kids	40%	24.4%	19.5%	16.1%	-5.7	312.1
	teens	48.8%	13.2%	24.1%	13.8%	-3.9	310.5
	m_teens	43.6%	21.8%	21.5%	13%	-4.4	146.6
	all	27.2%	16.5%	24.3%	32%	-3.6	273.7
w.c	kids	42.9%	18.1%	21.3%	17.5%	-6.5	253.6
	teens	39.6%	18.7%	21%	20.5%	-6.1	235.7
	m_teens	40.1%	19.5%	21.9%	17.5%	-5.77	259.8
	all	29.8%	18%	21.5%	30.5%	-3.6	239.3
m.r	kids	77.7%	4.3%	4.7%	13%	2.2	53.2
	teens	73.6%	3.9%	4.5%	17.9%	2.1	49.6
	m_teens	77.7%	4.5%	5%	12.7%	2.31	51.2
	all	59.2%	4.9%	6.4%	29.4%	2.1	69.2
p.q	kids	19.5%	24.6%	24.2%	31.5%	2.2	190.1
	teens	21.4%	23.1%	21.6%	33.8%	0.5	191.5
	m_teens	21.2%	25%	22.5%	31.2%	2	188.3
	all	13.6%	21.1%	18.9%	46.2%	1.3	182
s.c	kids	8.8%	29.9%	32.3%	28.9%	-2.6	52.2
	teens	9.9%	28.6%	32.4%	28.9%	-1.8	48.6
	m_teens	10.3%	31.2%	33.4%	24.9%	-2.38	52.8
	all	5.5%	23.9%	29.1%	41.4%	-1.6	48.6

Table 2.9: Click pattern frequency for the query reformulation types.

The same reformulation types were found to have the lowest ratio values in the young query sessions, which shows that these reformulation types are also preferred when users are not satisfied with their current queries. Nonetheless, a different behavior was found in the young query sessions since all the ratio values are higher than in the general purpose sessions. Thus, in the former sessions, the query reformulations are more frequently used to retrieve follow-up results. It is important to mention that the difference in the ratios are mostly caused by the significant higher usage of the pattern $click - click$ in the queries targeting content for young users.

This finding is aligned with Bilal [2001] studies on children search behavior. They observed that children need to explore further options before making a final relevance decision. This result also indicates that it is harder for users to find appropriate content for young users since it is necessary the exploration of more results

The ratio $\frac{(3)+(4)}{(1)+(2)}$ is an indicator of the effectiveness of the query reformulations. The most effective query reformulations were $w.r$, $w.c$ and $m.r$ and the least effective were wa, pq and sc in the children and general-purpose sessions. However the ratio values in the young user sessions were always higher given the greater number of clicks reported in these sessions.

The higher rate value of the $p.q$ query reformulation in the sessions targeting content for childrne are consistent with previous observations on children search behavior [Bilal, 2000]. in which children tended to repeat the same search even if the search did not return new results. This behavior is also in-line with their tendency to loop searches and links when solving complex information needs, as it was reported by Bilal [2001].

The mean rank change represents the rank position difference of the clicked domains in the reformulated and original query. For this analysis, higher rank values correspond to the results located in the bottom of the result list and lower values to the results on the top of the list. Thus, higher mean rank change indicates that the query reformulation is less successful since users click on results below the original queries. We found that all the query reformulations are successful in children and general-purpose sessions, except for $m.r$ and p.q. This result is logical since these actions do not involve query changes and users generally click on top ranked results first. It is interesting to note that although query reformulations are significantly less used in the sessions targeting content for young users (given the high used of the action $m.r$), these can be highly helpful given that the low values observed for the mean rank change.

The mean time change measures the average time in seconds that users wait to perform a query reformulation. It is measured by calculating the time difference in seconds between the sreformulated and original query. The wait time to perform query reformulations is longer for most types in the kids and teenager sessions. This result was expected since children's physical and cognitive skills are less developed

	All	Kids	F.Services
	Query level		
Query length	2.5%	3.8%	3%
Average rank	3.5%	5.7%	4.3%
Phrasal queries	56%	65.9%	42.3%
Question queries	2.7%	3.9%	2.1%
	Session level		
Session length	2.8%	8.6%	3.5%
Session duration	11.7%	20.2%	16.3%
	Query reformulation		
n.q.	48.1%	25.9%	49.2%
w.a.	1.5%	1.3%	1.6%
w.r.	0.1%	0.2%	0.2%
w.c.	6.4%	7.8%	7.1%
m.r.	38.8%	59.6%	36.9%
p.q.	3.0%	3.3%	3.1%
s.c.	1.8%	1.5%	1.6%

Table 2.10: Comparison of the data sets with the Finance Services directory.

than in grown ups. (e.g. type speed and reading skills), moreover children need more time to concrete their information needs and they are less focus during the search Bilal [2002]. Surprisingly the only reformulation type that doesn't follow this trend is the *m.r* type, which is the most frequently used in the children sessions.

2.8 Dmoz bias validation

It may be argued that the results presented above are biased since Dmoz users may have similar search behavior across the Dmoz categories. We show that this is not the case by replicating the analysis on an alternative Dmoz section. Particularly we employed the *Financial services* category since it targets a markedly different user group and its size is in the same order than the Dmoz Kids category. This category contains 14.181 Dmoz entries which lead to 261.605 query entries (38.593 unique) and 209.975 sessions using the extraction method described in this Chapter. Table 2.10 summarizes the results collected. From this table can be observed that although the query length and average click rank of the *Financial Services* data is greater than in the whole dataset, the values are still lower than in the users dataset targeting content for young users. On the other hand the percentages of phrasal and question queries are lower in the *Financial Services* data than the whole and young users datasets. Differences were also observed for the length and duration of the sessions. Additionally, a very similar behavior in the use of query reformulation types was

observed in the *Financial services* data and in the whole data set, although both differ from the young user sessions.

In general these results expose different search behavior in the users associated to the entire query log, young users and financial datasets, which indicates that the characteristics of the queries and sessions extracted using Dmoz differ according the domain of the information retrieved. This result is also consistent with White et al. [2009] studies in which it is shown that the search behavior varies according the domain of expertise of the users.

2.9 Conclusions

In regard to research question *RQ-1.1*, we found several differences with statistical significance in the queries and sessions employed to retrieve information for children, teenagers and general-purpose content. Thus, clear differences were observed in the search behavior of users targeting these types of content. To summarize, we found that the query structure and length of the queries in the four datasets varied. Longer queries were found in the queries targeting content for children and teenagers in respect to content for the average web user. Similarly, greater usage of natural language was found for the former query sets. The submission of longer queries shows the need of constructing specific and focused queries to access quality content for children and teenagers in state of the art search engines. At the session level we observed significantly longer sessions targeting content for young users. Similarly, we observed a higher number of clicks and result exploration, which did not imply the usage of more query reformulation in this type of sessions. The higher amount of clicks on these sessions was also reflected in a more disperse rank distribution and higher likelihood of clicking on lower ranked results in the sessions with young content clicks.

Some of our findings are in-line with previous studies of children's information-seeking behavior [Bilal, 2000, 2001, 2002; Druin et al., 2009]. The higher usage of natural language and in particular question queries is consistent with case-studies reporting children submitting question as queries for the case of complex information needs. Similarly, higher exploration of results and re-submission of the same query within the session are consistent results with the loopy search behavior observed in case studies with children under 12 years old.

Nonetheless, it is important to mention that the findings presented in this chapter address the search behavior of users searching for content for children and teenagers, and it is still not possible to confirm that the search behavior observed through this data is the actual behavior of young users on the web. For instance, we also observed evidence of parental supervision when analyzing the query types (e.g. larger usage of the focus query type). Understanding the search behavior of children requires further

research and a user centric approach with log data augmented with age-demographic information. This issue represents a limitation of this chapter that will be addressed in the following two chapters by using large-scale search and toolbar logs from users with reported age. However, the results presented in this chapter are valuable because they expose the difficulty that users have to find high quality content for young users. This was clearly reflected in most of the query log metrics employed in this study (e.g. longer queries, lower ranked clicks, longer sessions, loopy behavior). Moreover, the metrics explored in this chapter can be used to evaluate how state-of-the art search engines deliver high quality content for young users by providing an appropriate set of urls focused on this public. In this regard, it is worth to mention that we verified that the results presented in this chapter are not biased in the sense that all users may behave similarly across the Dmoz categories. We showed that the behavior of users searching for information related to finances behave differently from those users searching content for children, teenagers, and in general from the average web user.

In respect to *R.Q-1.2* we characterized that the intent of the queries collected is consistent with the intent reported in other studies for young users Ofcom [2010]. We also observed lower vocabulary sizes in the queries submitted by children and teenagers in respect to the average web user. From the cue word analysis carried out in Section 2.6.4, it was observed, from a qualitative point of view, that the information aspects (i.e. query senses) associated to the queries collected are representative of the targeted audience. In the next chapter we will compare from a quantitative point of view the topic distribution associated to the queries identified in this chapter (using the distribution derived from the topics of the Yahoo! directory) against the topics associated to queries submitted by young users.

In the following paragraphs we will provide recommendations and directions for future work.

2.9.1 Lessons learned and Recommendations

The better understanding of users retrieving information for children on a large-scale achieved in this chapter allow us to discuss several ways to improve the search experience of users searching for this type of content. We discuss improvements on two IR dimensions: query assistance and aggregated search.

Query assistance

The use of longer queries to retrieve children-friendly content shows that query expansion techniques can be highly beneficial to access this type of content, since these techniques allow the user to focus the search on specific information aspects. We observed in this chapter that query reformulations consisting of adding and changing tokens in the query lead to significant improvement in the ranking of the clicked

results. We believe, based on this evidence, that careful recommendation of queries for young users can make quality content for these users more accessible. In Chapter 5 we will explore query recommendation techniques for children and teenagers.

In this respect, cue words are also a valuable resource to rewrite queries by using the cues terms associated to the contexts words or to the entities that occur with the query. This method can ease the exploration of information by providing different content types and dimensions of the topic being searched. The study of further query rewriting techniques as the one presented in Zhang et al. [2007] can also be beneficial to reduce the cognitive load of reformulating queries.

Difficulty of finding high quality content for children

The low percentage of information queries found in queries targeting content for children suggests that although these users are familiar with interactive applications, they are not fully harvesting the information content available on the web.

We believe that there is an urgent need for more efficient ways to gather and present information to young users. Aggregated search is an adequate paradigm to address these difficulties. Aggregated search refers to the selection of results from diverse sources and content types, and the integration of this information to aid the user to reach their information need more efficiently Murdock and Lalmas [2008].

The identification and clustering of cue words shown in this chapter is a resource that be employed to assist the selection of relevant verticals for the information needs of young users. Although current IR systems for children such as *Yahoo Kids!* or *Ask Kids* already offer categories associated to some of the cue words that were identified in the clusters (e.g. games, coloring pages, poems, jokes, homework help, etc.), the dynamic selection of verticals is still limited. The detection of cue words from the query can be used as resource to complement vertical selection classifiers as the one presented in Arguello et al. [2009] by designing methods to match these cue words to the available verticals.

Additionally, state-of-the-art information retrieval systems only provide simple methods to aggregate results from the verticals. Further research is required to determine which verticals are relevant given the query of a young user, and to determine the best strategy to aggregate and simplify the content retrieved from the verticals.

In chapter 6 will be explore the creation of a test collection to study the problem of vertical selection for children. Vertical selection is a key step in the process of providing aggregated interfaces for children and teenagers. We will explore the usage of tags from social media as a source of information for the vertical selection problem.

Chapter 3

Analysis of Search Behavior of Young Users on the Web

This chapter is based on Duarte Torres and Weber [2011].

3.1 Introduction

In the previous chapter was analyzed the characteristics of the queries and search sessions of a set of queries targeting content for young users using the AOL search logs. The main limitations of the previous study are: *(1)* it is not possible to ensure that the queries were submitted by young users. *(2)* The queries and clicks analyzed represent the behavior of users accessing adequate content for young users, however this behavior can differ from the actual average behavior of young users.

In this chapter we address these limitations by exploring a large scale sample using Yahoo! Search logs with reported user age. Our focus is not only on characterizing their search behavior but on identifying the main difficulties that these users encounter when searching for information in the Internet.

The following two search sessions derived from this query log sample exemplified some of these difficulties:

(1) A 10 years old girl submits the query *what is love*, the search engine triggers advertisements related to dating and casual encounters. Thinking that this ad is a result to the query, the girl clicks on it, after spending few seconds trying to understand what is happening, the user goes back and then clicks on the first web result, which explains the chemical processes involved when people feel love. The content of this website goes beyond her reading skills and the user quits the search session, most likely dissatisfied; *(2)* When a 9 years old boy submits the query *hun*, the search engine suggests queries such as hun school (Princeton college), hun sen (primer minister of Cambodia) and hun empire (former empire ruled by Attila). Although

this user is probably targeting the last topic suggested by the search engine, the user does not seem to notice any of the query suggestions and simply continues with the initial query. Then, the user clicks on the first web result, which happens to be a web directory of links with adult content. Hun is also a popular term used to refer to a specific type of adult content. The user who is probably confused by the content, decides to go back, then the user clicks on the second web result, which is the Wikipedia article of the Hun empire. As it was the case with our previous example, the article is dense and its language is too advanced for the reading capabilities of this user, who after few seconds aborts the search session. In these two examples, we observed that young users have a tendency to click on higher ranked results, spend short times on each url, and in general have shorter sessions than those observed in older users.

We hypothesize that young users struggle in their way to find information on the Web (as it was the case with the two previous examples), and that these difficulties are quantifiable from their interactions with the search engine. Identifying the problems young users face when searching information on the Web, along their topics of interest is key in the development of search and web services tailored at children and teenagers.

The first aim of this chapter is to identify and quantify the search difficulties of young users using search engines and characterize the topics they are interested in, by using a large query log sample. We are also interested in contrasting the search difficulties and topic interests of young and adult users. At the best of our knowledge, little research has been carry out in this respect using search engines on a large scale. Even though several valuable case studies with child users have been reported [Bilal, 2000, 2001, 2002; Druin et al., 2009], these studies are highly obtrusive and include a limited number of users and a small number, which are assigned a limited number of artificial information tasks. These research studies are also unable to capture a representative overview of the broad spectrum of topics that motivate web search in the young population. In the same lime, we are interested in exploring signs of development stages through the usage of a search engine. For instance, to trace the evolution in terms of the readability of the content clicked and the sentiment reflected in the queries.

3.1.1 Research questions

The following research questions address the research aims expressed previously:

- **R.Q-2.1**: Do young users struggle to find information with a large scale search engine, and how is this struggle reflected in their search behavior from a query log perspective?

- **R.Q-2.2**: Does the search behavior and search difficulties of children, teenagers

and adults differ in a large scale search engine (Yahoo! Search)?

- **R.Q-2.3**: Can we retrace stages of children and teenagers development, in terms of the topics they are interested in, through their queries and the characteristics of these queries?

We tackle *RQ-2.1* to *RQ-2.3* by using a large query log sample from the Yahoo! search engine. The search logs registered queries and search activity of users from 7 to 70 years old. The logs were taken from the US market in a time window of 4 months in 2010. The methods employed in this study consist of using well established query log metrics and novel metrics targeted at understanding the search behavior and quantifying the search problems that children encounter during the search process.

In respect to *RQ-2.1* we hypothesize that young users (and particularly users up to 12 years old) have greater difficulties than adults to search for information on the web, and that this behavior is manifested in the way young users construct their queries and explore the result list. Particularly, we believe young users have problems specifying their information needs with keywords, and that young users make more use of natural language in their queries. In terms of result exploration, we expect young users to quit their search sessions sooner, and to have shorter clicks on results than adult as a consequence of their search frustration. We also expect to observe greater use of the same queries and visited results previously clicked within the search session, as it was reported by Bilal [2000, 2002]. Lower usage of query suggestion is expected since current search engines suggest terms for all user ages, thus most query suggestions are likely to be unsuitable for children.

We collect evidence to address these hypotheses through the following query log metrics: At a query level, we include average query length (measure of query specificity), query structure and natural language usage in queries. At a click level, we considered the click position and click duration distribution to analyze the way young users explore results. At the session level we explored session duration, session length, query re-finding, click re-finding, query suggestion and query correction usage. Further clarifications on how these metrics provide evidence of search difficulty will be given when reporting each result.

It is important to mention that after manually inspecting a small sample of search sessions, clicks on adult content and on advertisement at very young ages were observed. Motivated by this observation, we formally estimated the likelihood of accidentally click on adult content, and the likelihood of clicking on advertisement for young and adult users. These query log metrics are novel and oriented to collect evidence of the problems children encounter on the Web, since both type of content (adult content and ads) are clearly not targeted at children. High volumes of these clicks suggest search disorientation and difficulties to access appropriate information. We hypothesize that young children click on adult content by mistake more

often than adults, and that they have a comparable proportion of clicks on ads in respect to other age groups given their difficulty to identify this type of content as advertisement.

RQ-2.3 is addressed by pointing out differences in the topic distribution of the queries submitted by users of different ages. The topic distribution is extracted by classifying the queries submitted to the Yahoo! search engine using the Yahoo! Directory categories. We also employed a fine-grained topic classification using the categories of the Yahoo! Answers system, which allows to have a more detailed overview of the topics associated to the questions and concerns expressed in the children and teenagers queries. It is important to clarify that we are not analyzing the queries submitted to the Yahoo! Directory and Yahoo! Answers service, we solely used the category hierarchy provided by these two services to map the queries submitted to the Yahoo! search engine to topics.

We also show that the development stages are reflected in the search logs through the reading level of the pages clicked. As an aside, we employ other demographics factors such as the average income of the user's location to show its influence in the reading level.

The vocabulary size is another important feature of children development. Specifically, we measure the vocabulary size by counting the average number of unique words in the queries. We carry out an analogous procedure to find the *web resource* vocabulary of the users ,which is defined as the number of distinct *urls* and domains accessed by users in a given age range. As a last metric, the sentiment expressed in the queries is quantified. We expect greater usage of sentiment at younger ages.

We expect to find clear differences in terms of the topic distribution of children, teenagers and adults, and in general we expect to see a correlation between age and topic distribution. Similarly, we expect a higher proportion of clicks on basic language content and a smaller vocabulary size at younger ages given that children have less developed vocabulary and language capabilities. All the query metrics mentioned so far are also employed to address *RQ-2.2*. This is carried out by contrasting the results across age groups. For instance, we can characterize the way users submit queries by comparing the characteristics (e.g. query length, natural language usage) between users from 7 to 12 (children), 13 to 18 (teenagers) and over 18 (adults).

3.1.2 Chapter Organization

The chapter is organized as follows: in Section 3.2 we present the most relevant related work of previous studies on query logs and children search behavior. We also provide the most related literature of toolbar logs mining. In Section 3.3 is described the research method that is followed in this research.In Section 3.4 is presented the results and discussion of children's search difficult (although few of items shown in this

section also support *RQ-2.3*. Concretely we report results regarding query structure, click distribution, click duration, session characteristics and query assistance usage. Section 3.5 discusses our findings on retracing child development stages using query logs. These include topic distribution description, vocabulary size, query sentiment analysis and readability of the content clicked. Section 3.6 presents a comparison between the results obtained in the previous chapter using the AOL search logs and the findings collected in this chapter. Finally, section 3.7 concludes ths chapter with a discussion of our main findings and how they could be applied to state of the art search engines. We also provide recommendations for future work.

3.1.3 Limitations of this study

In this study, we focus on understanding the search and browse behavior of young users (children and teenagers). The findings are contrasted against the search behavior of adult users.

However, the analysis of the difference within adult users of different ages (e.g. young adults and seniors) is out of the scope of this study. Little research has been dedicated to understand the search behavior and difficulties that senior users encounter on the Internet. We believe an exclusive line of research is required to address this. The results presented in this study are derived from log data from the U.S. market in which the predominant language is English. Further research is needed to address cultural differences in search behavior and in pinpointing the search difficulties of users from different cultural origins and different languages.

The evidence collected throughout this study focus on search and browsing metrics captured by means of log data. Other aspects of the content explored by the users, such as the layout characteristics of the Websites, font types, sizes and other features that are not reflected in the query logs are not addressed and are out of the scope of this study. Further research is needed to relate aesthetic and functional features of websites with the search behavior of users.

3.2 Related work on query log analysis

Weber and Castillo [2010] presented a query logs study on how search differs on users with different demographics. They used demographic information derived from the US-census and user profile information to describe search patterns and behaviors for population segments with different demographic characteristics. In this chapter, we employed an analogous methodology to show that the reading level of the urls clicked by children also varies across demographic features.

Weber and Jaimes [2011] studied the relations between the dimensions "who searches", "what they search", and the "how they search" interact. Related to our

work, they also gave details about topical distributions as function of the user's age. Even though they provided detailed topic results for adult users, the young user topic characteristics are very broad since they aggregated the logs of users from 7 to 25 years old into a single group.

In our study we apply similar methodological techniques, such as analyzing session characteristic, however the main difference is one of breadth against depth, in which our focus is on a particular demographic group, namely young users (children and teenagers). In this way we provide more detailed results regarding the search behavior and difficulties that young users face when searching for information online.

Beheshti et al. [2010] performed an analysis of transaction logs from the portal *History trek*, which contains Canadian historic educational content for children. The logs analyzed contain up to 92K transactions and it was registered from 2007 to 2009. The authors found that the hierarchy of categories and subjects provided in the interface of the portal accounted for 83% of the searches, which indicates a clear preference for browsing over keyword based search. Beheshti et al. [2010] pointed out that this behavior is potentially due to the greater cognitive load associated to the formulation of search queries in respect to the browsing of predefined categories. Even though this result seems contradictory to some of the findings of Bilal [2000, 2002] and Fidel et al. [1999], it is important to recall that Beheshti et al. [2010] employed a carefully designed interface aimed at children, contrary to the first studies which were carried out on general purpose search engines.

More recent query log analysis have been carried out to understand the behavior of children when using search engines that have been adapted for them, in terms of search interface and content. Gossen et al. [2011] showed preliminary results from a query log analysis of three German search engines providing only content suitable for children. They found that children tend to submit shorter queries than adults (which is also found in our study) and that, as it is expected, children commit a greater number of spelling mistakes.

3.3 Method

We utilized the *TLA* method described in the previous chapter in Section 2.3 to analyze the sample extracted from the Yahoo! Search logs. The application of the *collection*, *preparation* and *analysis* steps are described in the following paragraphs.

3.3.1 Search logs data collection and Preparation

The data set employed in this study was extracted from a large sample of the Yahoo! Search logs from May to August of 2010. Only logs from the US market were included. The search interface that users employed correspond to the English Yahoo! US search

portal (`http://us.yahoo.com`), which is meant for all kind of audiences. No cultural differences were considered. In the data collection process, we only included log data from users for whom we could obtain (self-provided) age, gender and US ZIP code. We used the ZIP code in combination with the US census information[1] from the year 2000 in order to annotate users with demographic estimates about their education level (the fraction of the population in a certain age range holding a bachelor's degree or higher). This technique has been previously used in the context of query log analysis by Weber and Castillo [2010]; Weber and Jaimes [2011].

The search logs available to the authors consist of rows of event entries, each one associated to a time-stamp. Each event can refer to a user query submission or a click on a web result. In the latter case, the logs also provided the rank position of the clicked result. Each one of these entries was associated to the user profile information, which allows us to group log data from users within the same age group. The age groups are defined in Section 3.3.1.

The data was cleaned by first selecting log entries from users with a valid Yahoo! account. Thus, log entries of users with unspecified birth year, gender and zip code in the Yahoo! profiles were ignored. Log data from users with ill-defined fields were also excluded (e.g. invalid ZIP codes). This filtering step is compulsory in order to be able to identify the age of the user. The resulting data was cleaned further by applying the following criteria:

- Queries containing only a single token and that contain exclusively non alphanumerical characters
- Queries that were issued by only a single user within a given age group
- Queries containing personally identifiable information, such as credit card numbers or full street addresses

The first criteria was carried out by using a rule-based approach. For instance, regular expressions were designed to detect non-alphanumerical characters in the query string. For the second cleaning criteria we relied on the support of each query, which is obtained by counting the number of users that submit the query within a specific age range.

Only completely anonymous data was used in our study. All the user information that could lead to personal identification in the profiles were inaccessible. Additionally, all references to personal identifiable information (such as credit card numbers, street addresses, persona names) found in the queries were previously removed by Yahoo! and replaced with generic tokens. In this chapter, log entries with any type

[1] `http://factfinder.census.gov/`

of such generic tokens (from the data anonymization process) were disregarded, which refers to the third cleaning criteria.

The dataset obtained after applying these cleaning steps for users under 10 years old had a search volume in the order of hundreds of thousands of queries from tens of thousands of users. For users aged 10 or more years old we gathered search volumes in the order of millions of queries from hundreds of thousands of users[1]. As certain aspects of this data set are considered business sensitive, for various metrics we report *relative* differences between age groups, as opposed to *absolute* differences (i.e. actual click-through-rates on ads).

It is important to mention that a very large fraction of search volume originates from logged in users. Thus, the data analyzed is representative of the population targeted by the search engine used.

Data age segmentation

The motivation of this study is to characterize the search activity of young web searchers and identify crucial differences between the search behavior across children of different ages and adults. For this purpose, we aggregated entries from the log data according to the user's birth year. Concretely, we estimated the age of the users by setting the date of birth as the 31 of December of the birth year provided in the user profiles and considering that the search was carried out in 2010. The following age ranges were created:

- early elementary: 6-7 years old
- readers: 8-9 years old
- older children: 10-12 years old
- teenagers: 13-15 years old
- older teenagers: 16-18 years old
- young adults: 19-25 years old
- adults *(i)*: 25-30 years old
- adults *(ii)*: 30-40 years old
- adults *(iii)* : over 40 years old

[1]Exact statistics about the sizes of the datasets are not reported since they are considered business sensitive information

Our selection of the age groups follows the development changes present in these stages of life. Children from 6 to 7 years old refine their motor skills and start to be involved in social games. Children from 8 to 9 years old start to expand their vision of the world beyond their immediate surroundings. Children from 8 to 12 years old acquire the ability to represent the entities of the world in terms of concepts and abstract representations. Teenagers on the other hand become more interested in social interactions [Kail, 2009]. As we will show in this chapter, theses stages also have an impact on what children search on the Web, and on the way they interact with the search engine.

Arguably, the user information provided in the Yahoo! profiles is not trustworthy since people can lie about their age, gender or geographical location. Nonetheless, since at last, as early as 2007, Yahoo! has required the consent of a parent or legal responsible for users under 13 years old to create an account[1]. Currently Yahoo! charges a symbolic amount of $.50 to confirm that a guardian is responsible for the child creating the account. Apart from the (small) financial cost, the corresponding time and effort increases the chances of having veridical information for these age groups. Note that even if a small fraction of supposed child users lied about their true age, this is less problematic for general *trends*, though the actual *absolute* numbers will be affected. It is also interesting to notice that in social networks, children tend to lie to make themselves appear older, and this practice is often backed by parents [Richtel and Helft, 2011].

Note that the first three group of users map to the *kids* group defined in the previous chapter. The users aged 13-15 and 15-16 years old map to the *teenager* and *mature teenager* sets defined in the AOL search logs.

3.3.2 Search logs data analysis

In all of our work, we take a user-centric approach, as we want to provide insights into how young users search online. This means that *all* of our statistics are macro-averages, where metrics are averaged with each user contributing equally, as opposed to micro-averages, where metrics are averaged over all query instances and heavy users will have a bigger importance.

For various parts of the analysis presented in this chapter we employed the search session definition introduced in Section 2.4.

For the queries and clicked documents were computed various metrics, which will be explained in the sections where they are analyzed. However, the distinction between navigational and non-navigational queries [Broder, 2002] is used in several sections. We describe in this section how this distinction was computed. We combined two different approaches. First, we used the *click entropy* [Weber and Jaimes,

[1] `http://info.yahoo.com/privacy/us/yahoo/family/details.html`

2011] to get estimates about how diverse were the clicked results in response to a particular query. Queries that had a sufficient support, a minimum of two occurrences, were judged as navigational if the click entropy was no larger 1.0. This approach works well for head queries (e.g. detects the query "utube" as a navigational query). Additionally, we used a simple heuristic given (query, click) pairs. Note that this heuristic does not label the query as such as navigational, but rather individual (query, click) pairs. Therefore, *facebook* could be non-navigational if the user clicks `http://en.wikipedia.org/wiki/Facebook`. Our heuristics works as follows: First query and url are tokenized (by whitespaces and dot characters respectively), then tokens are sorted and plurals are stemmed. We label the pair as navigational if the query contains a domain extension (e.g., www, .com, .org), the domain of the url is contained in the query (or viceversa), or the edit distance between the query and the domain of the url is smaller that a threshold value. In the results reported we used two as threshold for queries containing more than four characters). For instance, this method is able to detect the navigational intent of the pair (kids abercrombi, `www.abercrombiekids.com/`).

Statistical Tests

In our arguments, comparisons between (macro-) averages computed for different groups, for instance session lengths of children between 6 and 7 and adults between 40 and 70, are core elements. Thus, we were careful to test the various differences for statistical significance. As it was done in the previous chapter, we employed the two-tailed t-test and we consider a difference to be statistically significant if the probability of the null hypothesis is smaller than 0.1%.

Most differences were also found statistically significant at a 0.1% level given the large volume of the data analyzed. For the non-statistical significant cases we will report p-values to provide a clear picture of the results reported. However, for simplicity we will omit reporting p-values that fall under 0.001.

3.4 Identifying and measuring search difficulty

Query, click and session characteristics were collected to identify differences in the search process between users of different ages and gender. In the following paragraph we analyze each one of these types of metrics. The focus is on finding metrics that give insight into the *search difficulty* that children face. In particular, we are interested in metrics related to *confusion*. These metrics addressed research questions *RQ-2.1* and *RQ-2.2*.

	Global		Non-Navigational		Navigational	
Age	# tokens	#chars.	# tokens	#chars	#tokens	#chars
6 to 7	2.55	16.49	2.80	17.40	2.14	15.16
8 to 9	2.56	16.59	2.77	17.27	2.24	15.91
10 to 12	2.56	16.62	2.81	17.54	2.17	15.49
13 to 15	2.60	16.82	2.84	17.67	2.18	15.67
16 to 18	2.64	17.08	2.86	17.92	2.19	15.68
19 to 25	2.71	17.34	3.03	18.72	2.22	15.16
26 to 30	2.68	17.34	3.02	18.83	2.25	15.32
31 to 40	2.65	17.43	3.00	18.88	2.26	15.76
> 40	2.80	19.05	3.12	20.09	2.43	17.73

Table 3.1: Query length averages by query intent.

3.4.1 Query length

Two query length metrics were considered: token and character length. Token length is measured as the number of tokens separated by white spaces, and character length is simply the number of characters (including white-spaces) in the query. Table 3.1 summarizes the results obtained by age range and query intent. A clear increasing trend was observed from younger to older ages. This result suggests that young users tend to formulate simpler information's goals compare to adult users. Given that the difference margin is larger for non-navigational queries, this result also indicates that young users have difficulties finding the right keywords to formulate more elaborated information needs. This result shows that for ambiguous topics is less likely that young users will retrieve the specific aspects that they are targeting, since the search engine provides content for all type of users, and the volume of information available for adults is significantly larger than the volume offered for young users, particularly children.

With respect to statistical significance, all the macro-averaged statistics were significant using the paired t-test (with p-values < 0.001), except for the following pair: token average for users aged 8 to 9 and users aged 10-12 (p-value= 0.064)

3.4.2 Natural language usage in queries

The aim of analyzing the usage of natural language in the queries is twofold: (1) As a mean to retrace children development. Children typically have a greater sense of curiosity [Cecchin, 1987], which we hypothesized is reflected in the searches they performed. For instance we expect a greater amount of question queries for users under 10 years old and greater usage of superlative constructs; (2) Children have been observed to pose queries in natural language given their lack of familiarity with the keyword approach of search engines. Greater usage of this type of queries provides

Age	quest.	modals	knowl. quest.	superl.	for kids
6 to 7	2.07%	0.41%	0.16%	0.91%	2.36%
8 to 9	2.56%	0.29%	0.08%	1.48%	1.74%
10 to 12	3.53%	0.58%	0.11%	1.46%	0.97%
13 to 15	3.84%	0.71%	0.16%	1.33%	0.43%
16 to 18	3.33%	0.69%	0.20%	1.15%	0.34%
19 to 25	2.80%	0.49%	0.20%	1.23%	0.32%
26 to 30	2.54%	0.44%	0.16%	1.16%	0.54%
31 to 40	2.19%	0.33%	0.14%	1.09%	0.68%
>40	1.69%	0.24%	0.11%	1.07%	0.31%

Table 3.2: Fraction of query types.

evidence of greater difficulty to express complex information needs through keywords, which are better suited for modern search engines.

We employed the same query types defined in Section 2.6.2 to quantify these phenomena.

Table 3.2 summarizes the results obtained by age for the set of non-navigational queries. On overall, we found that different construct types were preferred at different age ranges. For the case of question queries, we observed that the age ranges 10 to 12, 13 to 15 and 16 to 18 had the highest fraction of question queries, the 6 to 7 and 8 to 9 groups did not have a noticeably higher fraction than, for example, the 31 to 40 years age range. This result is surprising since in previous studies users under 10 years old have been observed to be more likely to submit this type of queries [Druin et al., 2009].

The "for kids" construct was prominent at the youngest group of users. This result may be an indication that there is self-awareness from this age group of the need of focused content *for children*. Another interpretation of this result is that this age group had greater supervision and aid from parents which are likely to add the *for kids* suffix to the queries. We believe this is an interesting result that requires further research to discern which interpretation is more accurate.

Lack of clear trends were observed for the other categories. Nonetheless, the fraction of superlative queries peaked for children in the 8 to 12 age range. This result was expected since this construct is one of the ways in which children of these ages express their curiosity about the objects that surround them.

We also aggregated the proportion of query types 1 to 4 into a single category for each age group. The ratio between this aggregate and the aggregate found for users over 40 years old was calculated. Figure 3.1 depicts the results. We observed that the greatest usage of natural language is found for teenagers aged 13 to 15, which are 2 times more likely to submit to a search engine this type of queries. High ratios were also observed for users aged 10 to 12 and 16 to 18. We expected to find this

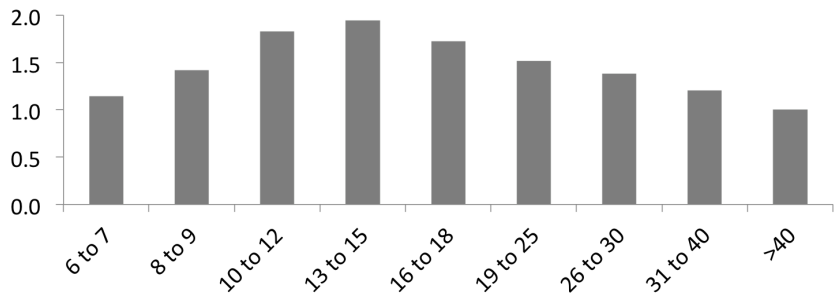


Figure 3.1: Relative frequency of natural language query types (1 to 4) of each age range against the group of users aged *>40*

behavior mostly on users up to 12 years old, which are the user segment that has been observed to use this type of construct the most [Bilal and Watson, 1998; Nahl and Harada, 1996]. Nonetheless, this behavior is also an indication of human development through the search queries. Teenagers are more prone than other user segments to make use of question answering systems such as Yahoo! Answers, in which a large number of adolescence topics are discussed by the participants (e.g. body changes during adolescence). We will elaborate on this finding in Section 3.5.1, in which we will address the topic interests for each age range.

When applying the paired t-test we found that the pairs 16-18 / 19-25 (*knowledge question*) and 16-18 / 19-25 (*for kids*) were the only cases in which the results were not statistical significant. The p-values obtained were 0.062 and 0.044 respectively.

3.4.3 Click position bias

We collect the macro-averaged distribution of the clicked result ranks. The macro rank distribution is computed by estimating the probability of each user to click at each rank position, only taking into account query instances with click, and then averaging the distributions across users belonging to the same age range. We set as cut-off 40 in the estimation of the distributions. That is, we accounted for all the click positions above the 40th ranked position. The motivation for this decision is to avoid spam. Figure 3.2 presents the distribution of clicks for the first 10 results. All click-through-rates are relative to those of users over 40 years old, that is the ratio between the proportion of clicks for the target rank position against the same proportion found for users over 40 years old. A rank of "0" refers to any kind of "special result" which includes links to current news, shopping results or any other high quality content, which is typically only shown for high volume queries. We allowed ourselves to exclude from this plot the relative click-through-rates of users between 19 to 40 years old to make the results displayed from young users (children and teenagers) more readable. On overall, we found that users over 18 years old behaved similarly in terms of click-

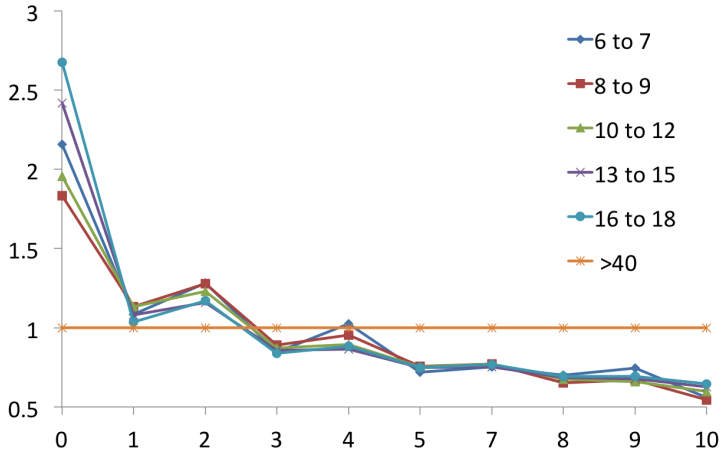


Figure 3.2: Relative rank frequency distribution across age ranges. The relative ranks (ratios) are estimated against the age group > 40.

through rates. On the other hand, the click distribution of teenagers (13 to 15, 16 to 18) and children (6-7, 8-9, 10-12) differed considerably.

Not surprisingly, we found that child users (6 to 12 years old) tended to click on higher ranked results, clicking twice as often as adults on the special rank 0 results. Interestingly, this behavior was even more pronounce for teenagers (13 to 18 years old), for which the ratio was around 2.50. It is important to mention that bias on top ranked results have already been observed in adult users. Höchstötter and Lewandowski [2009] reported that users prefer to click on links that are placed in the first position of the web result list. Craswell et al. [2008] proposed a framework to reduce this bias with log data.

Similarly, for positions 1 and 2 the ratios for children were between 1.10 and 1.30 times as likely as adults to click, while for teenagers the ratios were 1.10 and 1.15 respectively. For lower positions this behavior is reversed, that is, children and teenagers are less likely than adults to click below the third ranked position. On overall, we observed that children are less likely to click below the third rank position than teenagers.

These observations are consistent with previous studies reporting that children have a tendency to explore top ranked results and to avoid exploring thoroughly the list of results when they search for information. Fidel et al. [1999]; Schacter et al. [1998] reported this behavior and found that children felt frustrated and overwhelmed by the large amount of data returned by the search systems, and particularly by the lack of appropriate content. Our results support these observations, interestingly the same behavior accounted for teenagers. We will show that the frustration is also manifested in other search logs metrics such as click length and session duration.

The statistical tests were carried out based on the macro-average at each rank

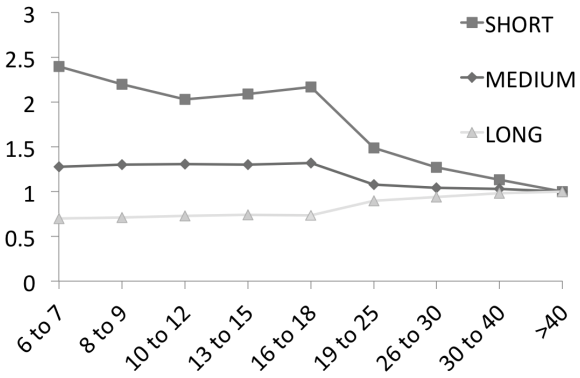


Figure 3.3: Distribution of click length across the age groups.

position. The following pairs were not found statistical significant: 6-7 / 8-9 (at rank 2 with p-value 0.364), 10-12/13-15, 13-15 / 16-19 (at rank 5 with p-value 0.103 and 0.091 respectively) and 8-9 /10-12 (at rank 7 with p-value 0.062). We did not carry out significance tests beyond rank 10.

3.4.4 Click duration

Previous work showed that a strong signal to detect search success occurs in the form of *long clicks*. Long clicks are defined as those clicks that last 100 seconds of more before another event is registered within the search session. Clicks at the end of a session are ignored from this analysis since they have unknown click duration. We broke down non-final clicks into the three classes suggested by Hassan et al. [2010b]: short (0-10 seconds), medium (11-99 seconds) and long ($\geq$ 100 second). Figure 3.3 shows that the fraction of long clicks is comparatively low for children of all ages, before it suddenly jumps to a higher level for users in the 19 to 25 age range. This result indicates search frustration in young users, since they tended to abort the clicked pages sooner than adults. For the short click macro averages, we found that all pair comparisons were statistical significant (p-values < 0.001). The following pairs were not statistical significant: 6-7 / 8-9 for medium and long clicks with p-values 0.142 and 0.287 respectively.

3.4.5 Click on ads

We employed the macro-fraction of ad-clicks to quantify how likely it is for a user of a given age range to click on an ad. Since not all the queries trigger advertisements, the estimation was performed only for clicks on results that were generated by queries that had triggered at least one click on an ad. Table 3.3 reports the fraction ratios of ad-clicks in respect to the group of adult users between 30 and 40 years old. Values

Age	Click ratio
6 to 7	1.28
8 to 9	1.14
10 to 12	1.04
13 to 15	0.89
16 to 18	0.90
19 to 25	0.84
26 to 30	0.80
31 to 40	1.00
>40	1.20

Table 3.3: Relative ad click through rates.

greater than *1* means that users were more likely to click on ads than the age range of adults between 30 and 40 years old.

Surprisingly, we observed higher ratios of ad-clicks for users at very young ages (6 to 12), which suggests disorientation during the search process for these users, since ads are, generally, not targeted at this demographic segment. This observation is in line with previous research that showed that in the context of online games children are also more likely to click on ads, as they fail to recognize them as such [An and Stern, 2011; Richtel, 2011]. It also reconfirms the findings concerning the position click bias from Section 3.4.3, given that we only consider ads that are displayed in the top area of the Website. Other advertisements slots are displayed in the bottom and side of the Websites. However, we did not find clear trends in terms on clicks for these two locations. All age group comparisons were statistical significant (p-values < 0.001).

3.4.6 Query assistance usage

Druin et al. [2009] reported in a detailed case study with 12 participants, that children aged 7 to 12 often ignore the auto-completion and query suggestion facilities provided by search engines. This behavior occurs because of their longer attention on the typing instead of the screen, which make children ignore the queries suggested by the search engine. Figure 3.4 shows the fraction of queries that were submitted to the search engine as a product of a query suggestion or query correction. Query suggestions are triggered by the search engine when the user is typing the query (e.g. query auto-completion) or as the form of related searches right after the user has submitted the query. The automatic query correction functionality is triggered by spelling mistakes and are commonly displayed by the search engine by informing the user. For instance: *We have included "britney spears" results - Show only "brittnay spears"*.

Figure 3.4 shows that children were more prone to use query corrections and

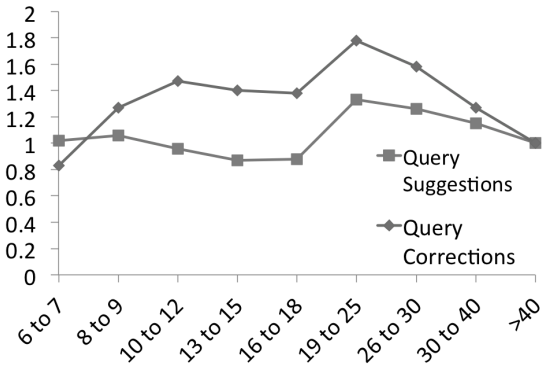


Figure 3.4: Query suggestions and correction usage.

that they were *not* more likely to make use of query suggestions. This result can be explained by the fact that query corrections are carried out by the search engine automatically, while query suggestions are only displayed and the user need to chose an appropriate query expansion, procedure that has a higher cognitive load. This behavior is inline with Druin et al. [2009] observations in regard to the the lack of focus experience when children are typing. However, the fact that query suggestions are even less used by teenage users (13 to 18), which have not been observed to have the same focus problems when typing, indicates that the suggestions provided by the search engine are simply not of interest for these users and consequently are not used. This is also the case for users about 40 years old. We also observed that the younger the users are, the more likely they are to *undo* query suggestions, or in other words, to insist on the incorrect spelling. For instance, by clicking on the option "Show only 'brittnay spears"'. The fraction of users aged 6 to 7 to click on such an in-correction was a factor of 1.62 higher than adult users. We believe that this behavior is a consequence of the lack of attention to the screen of young users when they are typing and the click bias. Recall that we found that urls located at the top are more frequently clicked by these users. To undo a query correction is necessary to click on the *undo* link, which is located at the top of the screen. The higher fraction of in-correction observed in young users explains partially why users aged 19 to 25 have a higher usage of this type of query assistance. However, more research is needed in this respect to understand why this particular age group makes more use of query corrections.

3.4.7 Accidental clicks on explicit content for adults

Children are potentially exposed to adult and explicit material on the Web given its large volume and the lack of parental supervision. Although, we observed lower volume of queries accessing adult explicit content for users under 13 years old (as it

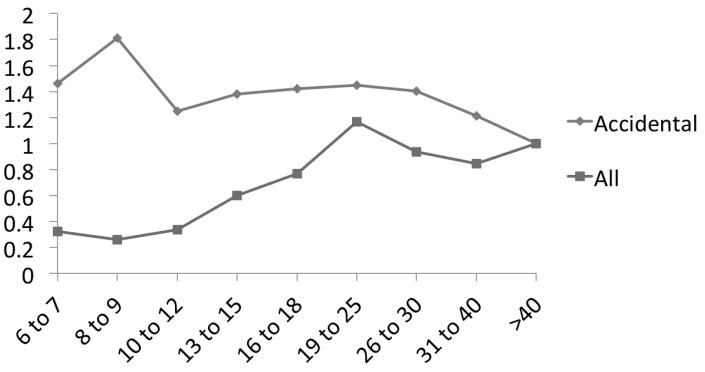


Figure 3.5: Relative likelihoods of accidental clicks on adult content websites. The *all* series refer to the relative frequency of clicking on adult content in respect to users over 40 years old.

will be depicted in Section 3.5.1) it is important to quantify how often this content is accessed accidentally.

We hypothesized that users clicking by accident on a website with adult content would immediately go back and click on a different web result. Thus, we estimate the likelihood of having a click on a website without adult content after a short click on a website with adult content, in the case when these events are registered during the same search session. Note that this process may occur more that once during the same search session. The last event of the sessions were ignored in the calculations since their click duration are unknown. Adult content was detected using a proprietary classifier based on the Yahoo! Directory[1]. Details about this topic classifier will be given in Section 3.5.1.

Figure 3.5 shows the relative frequencies for the event of an accidental short and immediately reverted click on adult content. This figure also shows the relative frequency of clicking on adult content in respect to users over 40 years old. On overall, the click on this content increases from the youngest group of users and picks at users from 16 to 19 years old. Note that even though children in the 6 to 9 years age range have a comparatively high probability of immediately reverting to a different result, after a (supposedly accidental) click on adult content, their *absolute* probability of clicking on this type of result or of issuing a related query is very low.

The fact that the probability of these accidents-with-immediate correction are higher for children aged 6 to 7 than for children aged 8 to 9 can potentially be explained by the fact that the youngest children might take longer to read an entry page explaining that the site contains adult material and that the visitor needs to be of legal age (typically 18) to view the content. All the age group pairs were statistical

[1] `http://dir.yahoo.com/`

Age	S.duration	S. length	Query ref.	Click ref.
6 to 7	3.79	3.76	0.24	0.17
8 to 9	3.51	3.71	0.24	0.15
10 to 12	3.63	3.71	0.23	0.14
13 to 15	3.91	3.76	0.26	0.14
16 to 18	4.04	3.82	0.26	0.13
19 to 25	8.20	5.45	0.32	0.22
26 to 30	8.45	5.43	0.30	0.20
31 to 40	8.39	5.28	0.29	0.20
>40	8.42	5.25	0.34	0.24

Table 3.4: Session characteristics.

significant and all the paired t-test p-values were very small (< 0.001).

3.4.8 Session characteristics

A sign of search confusion occurs when a user goes back to a query issued earlier in the same session after temporarily exploring different queries. As our sessions were quite short, with an average of 3.51 minutes for users between 8 and 9 years old, it is unlikely that the second occurrence of a query indicates renewal of the earlier information need. More likely, it indicates that the user has not yet fulfilled the earlier information need. We call those queries that are repeated within a session "query refindings", and their fraction is computed as follows. For each user we estimated the fraction of refinding queries inside a session (in respect to the total number of queries inside the session). Then, we averaged these fractions for all the sessions of the same user to generate a per user estimate of query refinding usage. As with all the other metrics reported in this chapter, we reported the macro-average across users. Similarly, a user clicking the same url repeatedly (which can be interspersed with other events) can be seen as an indication that the user is struggling and trying to make up their mind about the most relevant result.

Table 3.4 shows the fraction of refinding queries and clicks. This table also shows two simple measures for the average session length: the length measured in minutes and number of events (queries, clicks and next result page) in a session. It is important to clarify that these estimations excluded sessions containing only one entry (i.e. sessions in which a query was submitted and no clicks were registered).

Table 3.4 shows that the search sessions of children are considerably shorter that the sessions of adults. Surprisingly, this also includes the 16 to 19 age range and the jump to "adulthood" occurs suddenly in the group of 20 to 25 years old. This result suggests a greater level of frustration of young users since they tended to quit the search session earlier. This observation is supported by the fact that young users (under 19 years old) had a significant higher proportion of short clicks, as it was

pointed out in Section3.4.4.

The fraction of query refinding and, in particular, click refinding is *lower* for children. However, rather than to be taken as an indication of a lower level of confusion, this observation is more likely to be due to the fact that children have (considerably) shorter sessions, thus there is simply less opportunity to issue the same query or click.

Non-statistical results (at 0.01%) were found for the following cases:

- Macro average query refinding: 6-7 / 8-9 (p-value 0.231), 13-15 /16-18 (p-values 0.016)
- Macro average click refinding: 10-12 / 13-15 (p-value 0.061)
- Macro average session length: 8-9 / 10-12 (p-value 0.053)

3.5 Tracing children development stages

Previous work showed that, given enough search user history, attributes such as gender, age and location can be estimated Jones et al. [2007]. In this chapter, we looked at a related but different problem: can we find hints in the query logs that give indications about a child's development stage? (*R.Q-2.3*)

We hypothesize that human development can be traced through the topics and entities targeted by the queries, the gender topic difference, the sentiment expressed in their queries, the reading level of the content clicked and the query vocabulary. In the following sections we explore this hypothesis in detail.

3.5.1 Topic distribution

We investigated what children search for and how the searches evolve along two dimensions: topics and entities. For the first, we employed a high level classification of topics based on the Yahoo! search directory and a novel fine-grained classification of topics based on the Yahoo! Answers service. For the second, we explored the concrete entities they search and the characteristics of these entities.

High level topic distribution

We used a proprietary classifier to map web pages to entries of the Yahoo! Directory[1]. We used a weighted majority voting scheme on the top 10 organic results returned by the Yahoo! search engine to obtain a classification for *queries*, instead of pages. Weber and Jaimes [2011] provides details of this classification method. In total, there were 95 topics. Figure 3.6 presents the average topic fractions for the 11 most frequent

[1] `http://dir.yahoo.com/`

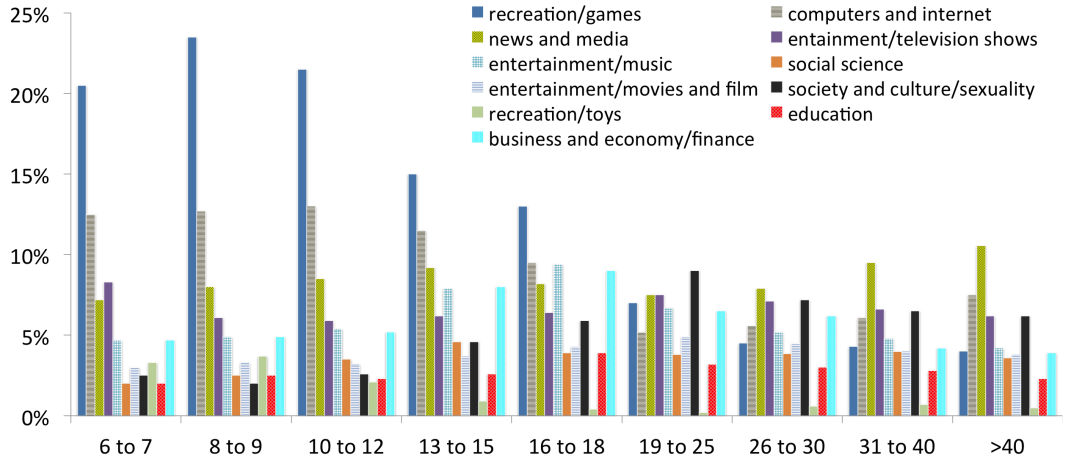


Figure 3.6: Topic progression through the ages.

topics searched by users in each age range. Note that we are using queries submitted to the Yahoo! search engine. The Yahoo! Directory is only used to map these queries to categories.

The behavior of Figure 3.6 is intuitive. Children up to 12 years old had a much higher fraction of queries falling into *recreation/games* than adults and the same holds, at a lower extent, for *recreation/toys*. The interest in music is mostly expressed in the teenager age ranges (13 to 18). The fraction of *business/finance* increases steadily for older users. We also observed a higher diversity of topics for users over 19 years old. In other words, few topics represent a large volume of queries for users under 19 years old, while the queries of older users are more evenly distributed in a larger number of topics.

Even though, we are most interested in understanding age-related differences, there are also important gender-related differences, even in children [Eccles et al., 1993, 1990]. We were interested in how gender differences evolve as children grow up. We were wondering if gender differences more pronounced in, for example, teenagers than in adults. To answer this question we quantified gender differences by looking at the topical distribution for particular age groups. Each one of such topical distribution corresponds to a probability distribution, summing to 100%. We used the 1-norm to quantify the differences between the probability distributions of males and females for each age group. The 1-norm is estimated by summing up the topic proportion values within the topic distribution of each age group[1].

The blue line in Figure 3.7 shows that the gender differences for children are significantly smaller than for adults. However, many of these gender differences are due to a gender bias introduced by adult content topic. The red line shows the

[1]http://mathworld.wolfram.com/L1-Norm.html

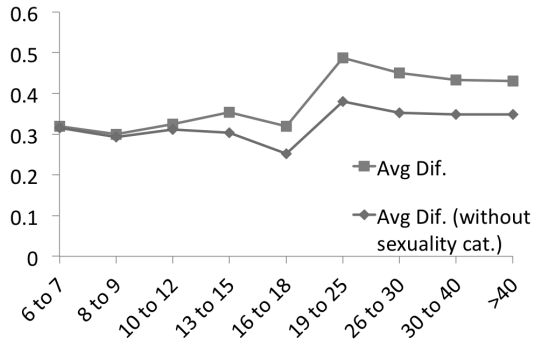


Figure 3.7: Average topic difference between genders through the ages as measured by the ||1||-norm.

gender differences when this topic has been removed and the remaining topics have been renormalized. As can be seen in the plot, this modification removes a large part of the age-related difference between the genders.

The largest differences between genders were observed in the categories *business and economy*, *computers and Internet* and *society - culture/sexuality*. Nonetheless, these differences were significantly higher for males and females over 16 years old, which is the trend that was observed in Figure 3.7.

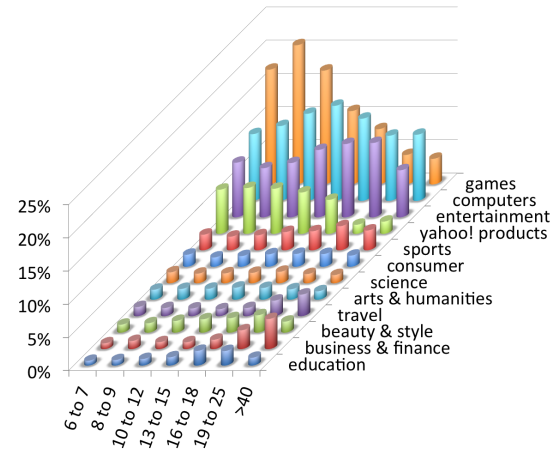
Statistical significance was tested by comparing each topic percentage across age ranges (using the paired t-test at 0.1% level of significance). All results were statistical significant with very small p-values (< 0.001).

Fine-grained level topic distribution

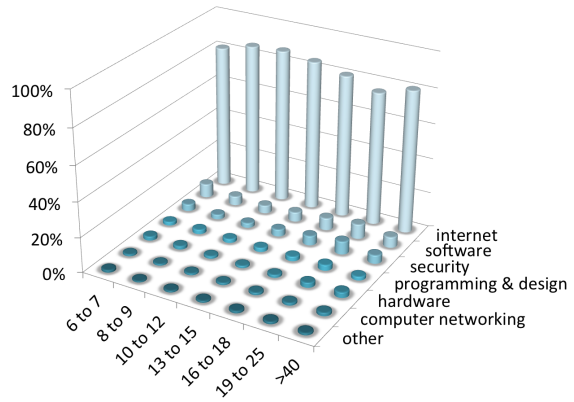
In this section, we employed the category structure of the Yahoo! Answers service[1] to have a more detailed classification of the topics targeted by the queries. Concretely, the classification is carried out by submitting the query to the Yahoo! Answers system and the majority vote scheme is used on the categories associated to the top *10* answers for the query. As it was the case with our previous analysis, the Yahoo! Answers system is only employed to classify the queries using the topic hierarchy, recall that the queries studied were originally submitted to the Yahoo! Search engine.

For our analysis, we employed the main categories (e.g. *games & recreation*) and subcategories at depth 1 (e.g. *games & recreation*, *video & online games*). This classification gives us a better overview of the most important topics associated to the concerns and queries that arise in users of different ages. Pinpointing the most frequent topics associated to this type of queries can help designers of search engines for young audiences to focus on certain topic that are critical for these users. For in-

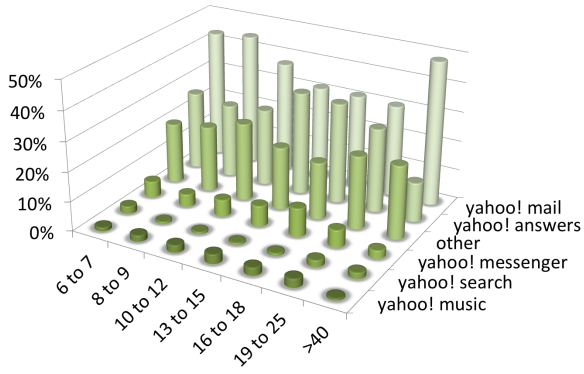
[1] http://answers.yahoo.com/



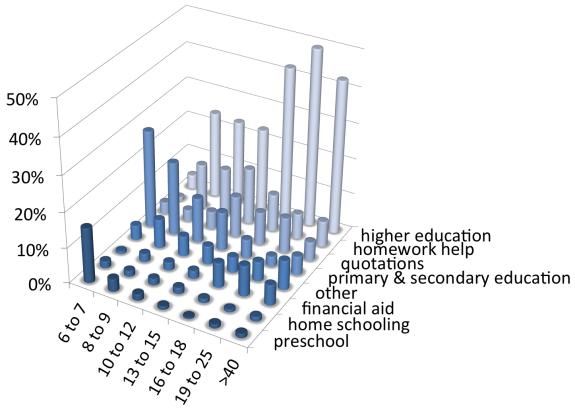
(a) Global



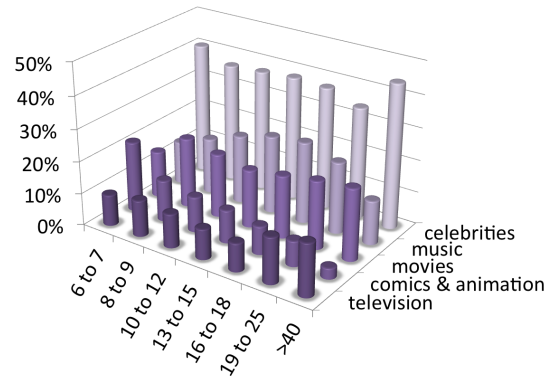
(b) Computers



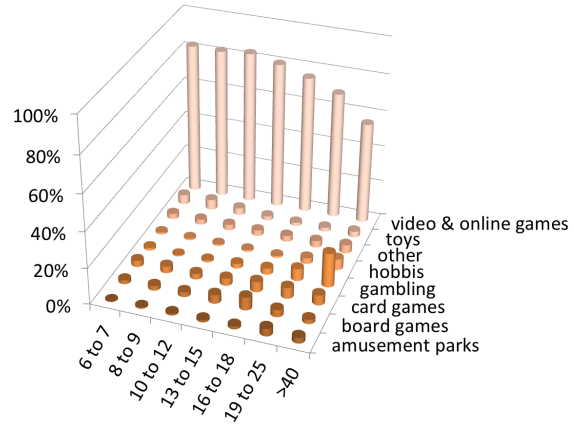
(c) Yahoo! products



(d) Education



(e) Entertainment



(f) Games

Figure 3.8: Distribution of topics for informational queries

stance, the distribution of topics can be employed to select the most relevant verticals in a search engine according the age of the user.

Concretely, the set of informational queries and a large sample of *how-to* queries were employed in our analysis. *How-to* queries provide a cleaner picture of the concerns that arise in young users. These queries were identified by matching informational queries with the prefix *how-to*. Although more sophisticated mechanisms to identify *how-to* queries have been addressed in previous research [Weber et al., 2012], we believe that this approach provides results with high precision. Given the rapid development of children, we expected to see a rapid change of the topic distribution of the *how-to* queries given that their interests and development stages evolve constantly.

Figure 3.8(a) depicts the distribution of topics for all informational queries using the global categories. In total, up to 24 general categories and 140 subcategories were identified by mapping the query set to the Yahoo! Answer categories. We observed a strong long tail effect in the topic distribution of queries of adult users. For instance, non-frequent topics (topics that account for less than 5% of the total volume of queries) sum up to 20% of the volume for each one of the groups of users under 12 years old, up to 25% for users up to 19 years old and 38% for users over 19 years old. These observations indicate that the queries of young users are concentrated of a fewer number of topics, while queries submitted by adults are more diverse. This is also reflected in the vocabulary size of young and adult users, as we will show in the next section.

We found that the dominant topics (in increasing order) for users up to 12 years old were: *games & recreation*, *computers & internet*, *entertainment & products*, and *sports*. For teenagers, we observed the same set of dominant topics and *computers & internet*, *entertainment & products*, which had a greater volume of queries than *games & recreation*. This was also the case for users over 19 years old, although the categories *business & finance* and *travel* also gained importance for these users.

We broke down these topics into further detail using the subcategories of the Yahoo! Answers directory. Figures 3.8(b) to 3.8(f) depict the results obtained for the subcategories with the highest volume, and the subcategories for which there are significant differences between ages. For *entertainment & products* we found that although the query volume of this topic is large for all the age ranges, the specific *entertainment* subcategory that is searched varies greatly according the age of the user. For instance *comics and animations* is searched 6.5 times more by users aged 6 to 7 than by adults, and up to 2.5 more than teenagers. *Music* is searched 1.9 more by teenagers in respect to children and 1.5 times more in respect to adults. On the other hand, *celebrities* are searched evenly by all the age groups. Figure 3.8(e) shows the entertainment subtopics distribution for all the age ranges.

For the *yahoo products* category we observed that most of the search volume fall

under *email* and *yahoo answers*. In particular, we observed that users up to 12 years old and over 40 years old target 1.5 more frequently queries to access email services than the others age groups. We also observed that queries from users up to 12 years old fall under *Yahoo! Answers* twice as often as adults while teenagers had the greatest percentage of queries for this category (2.6 times the amount of queries than adults submit under this category). These observations indicate that the Yahoo! Answers service is a highly valuable resource for young users, particularly teenagers.

For *Computers & Internet* most of the volume was found under *internet*, *software* and *internet security* for the case of adult users. *Internet* is an umbrella for topics related to search engine usage, social networking websites and popular encyclopaedia services as Wikipedia. *Software* is related to the installation, usage, and general support of all kind of computer software. For the exception of *security*, for which the volume of queries for adult users is twice the volume of the other groups, we did not observe clear dissimilarities in the distribution of these subtopics between the age groups. For the case of the games' categories most of the queries were related to *video games* for all the age ranges, as it is illustrated in Figure 3.8(f).

The results described so far show clear differences and changes in the topic interests of users of different ages. We also quantified objectively the difference between the topic distributions and the difference for each topic between children and adults. The differences between topic distributions were measured using the correlation between the topic distribution of a given age range and the topic distribution of users over 40 years old. The Pearson's correlation coefficient was employed for this purpose. In our context, a positive correlation means that users for the target age range tend to submit queries with the same topics queried by users over 40 years old. On the other hand, lower correlation values mean that the users of the target age range submit queries with different topics that the ones queried by users over 40 years old, which means that is harder to differentiate between both group of users based, solely, on their query topic distribution.

The difference within topics was measured using the Pearson's correlation of each topic and the age of the user. Specifically we estimate the Pearson's correlation for each topic using as variables the lower age in each age group (e.g. 6 for the age group of 6 to 7) and the proportion of the target topic in the given age group. In this case a negative correlation means that the topic is more prone to be used by younger users. A positive correlation means that the topic is more frequently queried by older users. A similar approach has been employed to measure the *trendiness* of words in scientific publications [Hiemstra et al., 2007].

Figure 3.9 presents the correlation between the topic distributions for the set of informational and how-to queries. This figure shows a marked trend, in which higher correlation values are obtained the older the user is. These observations indicate a

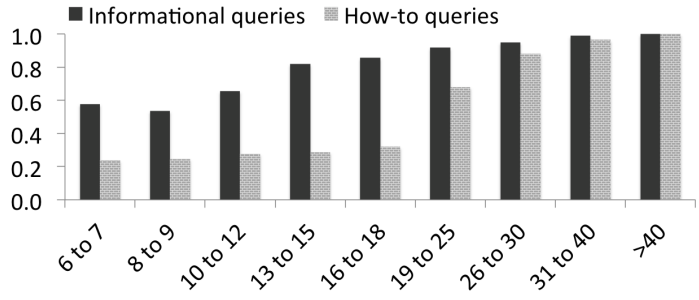


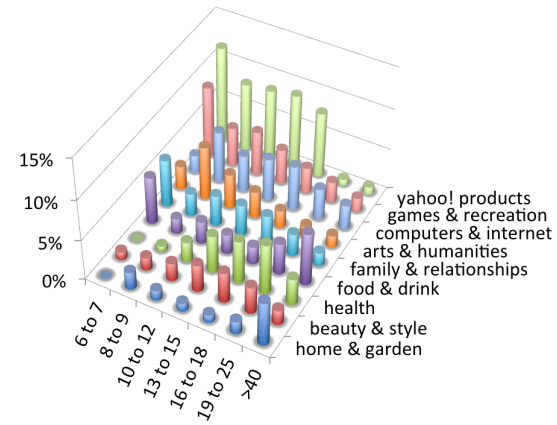
Figure 3.9: Pearson's correlation of the topic distribution of each age against the topic distribution of users over 40 years old.

Topic	P. Correlation
yahoo! products	-0.80
games	-0.78
arts	-0.53
pets	-0.40
science & math.	-0.40
computers	-0.15
entertain.	0.09
education & ref.	0.34
beauty & style	0.44
consumer electronics	0.56
sports	0.70
health	0.84
family & rel.	0.86
travel	0.93
business & finance	0.95
news & events	0.96

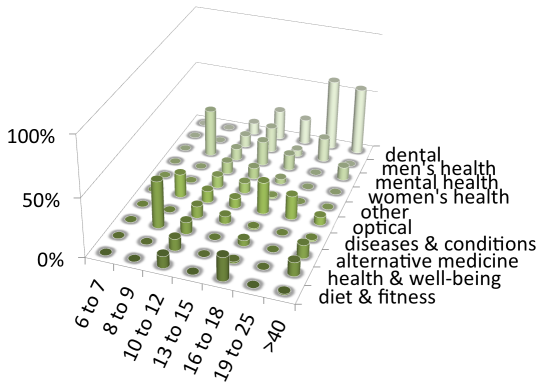
Table 3.5: Pearson's correlation of informational query topics and the age of the user.

clear distinction between the topics queried by young users and adults. Interestingly, we observed that the correlation gap between young and adult age groups is bigger for the how-to queries.

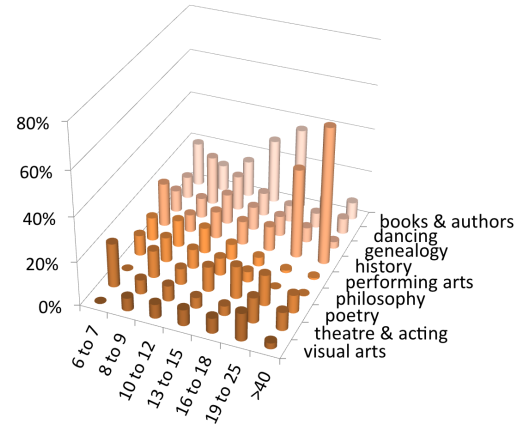
Table 3.5 presents the correlation within topics and the age of the user for the set of informational queries. We found that the topics *Yahoo! products*, *games*, *arts* and *science & mathematics* are highly correlated with young ages. On the other hand, topics such as *news*, *business & finance* and *travel* are highly correlated with older ages. Categories such as *computers & internet* or *entertainment* did not have a clear correlation with young or old users.



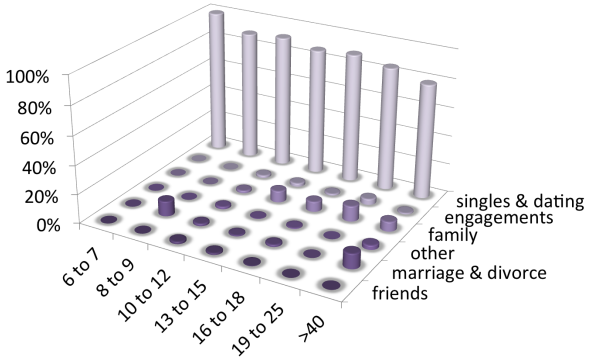
(a) Global



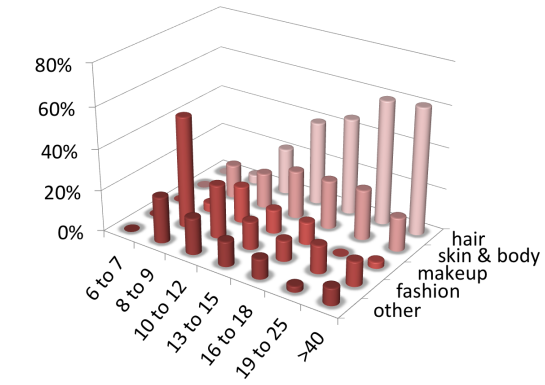
(b) Health



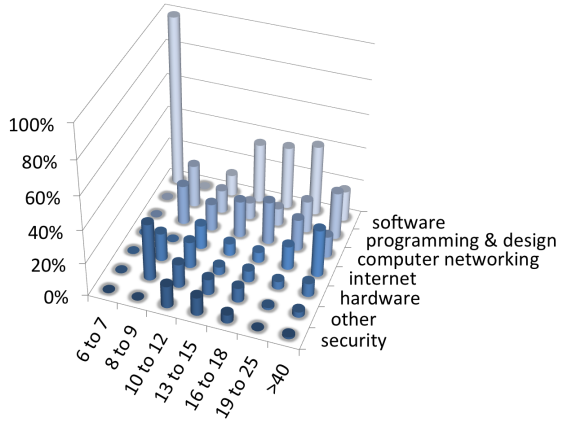
(c) Art



(d) Family



(e) Beauty & Fashion



(f) Computers

Figure 3.10: Distribution of topics for how to queries

Specific topics children target with how-to queries

Figure 3.10(a) presents the topic distribution obtained with the set of *how-to* queries. As it was the case with the set of informational queries, we found that *yahoo products*, *games* and *computers* represent a large percentage of the volume of queries of young users, with 22%, 14% and 10% of the query volume for users up to 12 years old. The percentages for users between 13 and 18 years old were 18.31% 8.01%, and 11.60% respectively. We also observed a large volume of other topics such as *art & humanities* (10% for users under 10 years old and 9.6% for users between 12 and 18) and *family & relationships* (5.7% for users up to 12 years old and 8% for users between 12 and 18 years old). *Health* and *Beauty* were also prominent for teenagers (10.6% and 7.61% respectively). For adults, the categories *food & drinks* and *home gardening* (12% and 6.7% respectively) had greater importance than for young users.

We also observed the long tail effect difference observed in the previous section between the young and adult group of users. Specifically *non-frequent topics* sum up to 30% of the query volume for users up to 12 years old and around 36% for users over 12 years old. A particular case was the age range of 6 to 7 years old, in which only 8% of the query was associated to *non-frequent topics*.

As we did in the previous section, we explored the subcategories associated to the *how-to* queries. For the subcategories of *family & relationships* we found that most of the queries targeted topics related to *singles & dating*. Surprisingly this was also the case for the youngest group of users (90%, 80% and 78% for users up to 12, 13-18 and over 18 years old respectively). The queries of the youngest group of users classified under this subcategory were manually explored. It was observed that these queries were often regarding early curiosity on dating. Some exemplary queries of this subcategory are: *how to kiss a girl first time*, *how to know what she wants from me? how to find out he likes me?*

Marriage & Divorce was particularly prominent for users over 40 years old (9%). Interestingly, this was also observed for users aged 8 to 9 years old (10%), which are old enough to understand this concept and at the same time being affected by it. For the children group we found that queries falling under this subcategory often reflected concerns regarding parents' relationships (e.g. *how to know if mom work things out with dad*). The subcategory *family* also had a high volume of queries for the case of users over 18 years old (12%).

For the topics related to *art & humanities*, we did not observe a clear dominant subcategory within age groups. In fact, the volume of queries under this category is distributed among most of the subcategories. The distribution of subcategories is shown in Figure 3.10(c). Nonetheless, we observed large differences among the subcategories across age groups. For instance, queries targeting topics related to the subcategories *books & authors* and *dancing* were up to 3 times more frequent for users

Topic	P. Correlation
yahoo! products	-0.82
family & rel.	-0.77
games	-0.75
arts	-0.65
entertain.	-0.52
computers	-0.44
society & culture	-0.21
sports	-0.03
beauty & style	-0.01
consumer electronics	0.00
pets	0.04
health	0.36
food & drink	0.67
business & finance	0.85
home & garden	0.92

Table 3.6: Pearson's correlation of how-to query topics and the age of the user.

up to 12 years old than for adults. The gap between these two sets of users was even greater for the subcategories *performing arts* and surprisingly for *philosophy* (up to 8 times more frequent than adults). Examples of queries under *philosophy* are: *how to take back words? how do you define good and evil*. *History* was more prominent for adults, accounting for 53% of the query volume under this category (4 times more than in the children and teenager ages)

For the category *beauty & fashion* we found that *fashion and accessories* was the dominant topic for users up to 12 years old, accounting on average for 53% of the search volume, which is 3 times the volume found for adults. The dominant subcategory for teenagers and adults was *hair* with 55% of the search volume (3 times higher that the volume accounted by users up to 12 years old). It was also observed that *makeup* and *skin and body* were prominent for users between 12 and 19 years old (10 times higher the number of searches in respect to adult users). Further details are shown in Figure 3.10(e).

Finally, for the categories *yahoo! products* and *computers* were found similar trends as the ones reported for the set of non-navigational queries. The results for the *computers* category are depicted in Figure 3.10(f).

As it was the case with the set of non-navigational queries, we measured the correlation with the topics extracted with the how-to queries and the age of the users. Table 3.6 depicts the results obtained. We found that the topics *yahoo! products*, *games*, *arts* correlate negatively with age. On the other hand, the topics *business & finance*, *health* have a positive correlation with age. This was also the case with the set of non-navigational queries. Interestingly, we found that how-to

Query	Entity
facebook, facebook login	`en.wikipedia.org/wiki/Facebook`
disney cars games, 2011 cars, cars 2 2011	`en.wikipedia.org/wiki/Cars_2`
what to do with hummus, ideal protein	`en.wikipedia.org/wiki/Hummus`
youtuyoutube, youtui, youtuyoutube	`en.wikipedia.org/wiki/Youtube`

Table 3.7: Examples of queries and their mapped entities.

queries categories such as *family & relationships* or *entertainment* have a negative correlation with age. This result shows that the characteristics of the query (e.g. how-to queries) can influence the topics that correlate best with young users.

3.5.2 Entity targeted by the users' queries

As the topics we used were fairly broad, such as "music" or "finance & investment", we were also interested in obtaining more fine-grained information by looking at the (main) *Wikipedia entity* that a query refers to. To map queries to Wikipedia articles we used the following simple, yet effective, approach: we sent the queries to the Yahoo! search engine and limited the results to results from `http://en.wikipedia.org/wiki/`. The first result was used as the entity representation for the query. Note that the queries sent to Wikipedia were run fairly recently, though the original queries were submitted about one year earlier. This ensured that even for recent events almost always a Wikipedia page could be found. Table 3.7 shows some examples of this mapping.

An overview of the entities searched by young and adults users is presented by the tag clouds in Figure 3.11 and 3.12 respectively. These entities correspond only to non-navigational queries. Entities related to adult content were manually removed.

One of the advantages of mapping queries to Wikipedia pages is that Wikipedia pages come with a categorical classification and that this classification is both more fine-grained, and in a certain sense orthogonal to our own topic classification (see Section 3.5.1). For example, pages about current celebrities almost always belong to the "Living People" category. Similarly, there are many child-related categories such as "Early childhood education". We used a simple pattern match for the prefixes "child" and "kid" to identify these pages.

In Table 3.8 we present the fraction of entities associated to children content in Wikipedia and the fraction of famous people found in the queries for the age groups.

Given that the celebrities entities often refer to trendy artists, which are more known by teenagers, we expected children and teenagers to have a large fraction of celebrity entities. However, that did not turn out to be the case and the highest fraction of such queries was observed for older user, particularly users in the range

Figure 3.11: Entity tag cloud: 10 to 12 years old.

Figure 3.12: Entity tag cloud: over 40 years old.

Age	Children and Kids	Living people
6 to 7	5.81%	8.11%
8 to 9	5.49%	6.97%
10 to 12	3.38%	7.59%
13 to 15	1.46%	9.47%
16 to 18	0.95%	10.86%
19 to 25	0.62%	11.96%
26 to 30	0.63%	11.54%
31 to 40	0.89%	11.09%
>40	0.62%	10.88%

Table 3.8: Entity fractions for Child related content and Living People according to the Wikipedia categories.

Age	Positive	Negative	Diff
6 to 7	1.233	-1.211	0.0216
8 to 9	1.253	-1.237	0.0161
10 to 12	1.257	-1.248	0.009
13 to 15	1.284	-1.274	0.0101
16 to 18	1.274	-1.258	0.0165
19 to 25	1.300	-1.283	0.023
26 to 30	1.302	-1.275	0.026
31 to 40	1.322	-1.297	0.0248
> 40	1.400	-1.376	0.0279

Table 3.9: Mean query sentiment values scores.

19 to 25 years old. The trend for child-related categories behaved as expected given that the fractions are more pronounced the younger the user is.

All proportions shown in Table 3.8 were statistical significant using the paired t-test between pairs age range proportions (all the p-values found were < 0.001).

3.5.3 Sentiment expressed in queries

Kuhlthau [1991b, 1999] found that uncertainty during the search process, unfamiliarity with technologies, and lack of ownership of the search task (e.g. when children are asked to search a topic by their teacher) can lead to anxiety and frustration, which we hypothesize is manifested in the sentiment expressed in the users' queries. Additionally, children aged 9 to 12 years old have been observed to experience extreme changes of mood [Smith et al., 2011], which we hypothesize may also be reflected in the formulation of queries. Chelaru et al. [2012] also attempted to measure the sentiment of queries by employing a set of queries targeting controversial topics.

To assign numerical scores to the sentiment expressed in the queries, we used the SentiStrength[1] tool developed by Thelwall et al. [2010]. This tool simultaneously assigns both a positive and a negative score to fragments of English text, since users can express both types of sentiments at the same time, such as in "I love you but I also hate you". Positive sentiment strength scores range from $+1$ (somewhat positive) to $+5$ (extremely positive). Similarly, negative sentiment strength scores range from -1 to -5. The tool works by assigning scores to tokens in a dictionary, which includes common emoticons. For example, "love" is mapped to $+3/-1$ and "stink" is mapped to $+1/-3$. Modifier words or symbols can boost the score such that "really love" is mapped to $+4/-1$ (the same for "love!!" or "looove"). The final positive sentiment strength for a bit of text is then computed by taking the maximum score among all individual positive scores. The negative sentiment strength is similarly calculated.

[1] `http://sentistrength.wlv.ac.uk/`

As can be seen form Table 3.9, the sentiment analysis applied to individual queries did *not* reveal the expected trend. It did however reveal that the tendency to use both more positive and negative words in a query *increases* as users get older. This phenomenon is at least partly explained by the fact that they issue longer queries (see Table 3.1) and hence the probability of positive/negative sentiment words appearing is higher. We observed the same behavior when measuring the sentiment of the set of non-navigational queries.

Further research is required to explore our hypothesis in respect to the expression of sentiment through the queries. Current tools to measure text sentiment are not designed for very short pieces of text, which is the case of search queries. The development of more suitable tools to measure sentiment in short texts would be beneficial for this research along the quantification of the sentiment in the content clicked by the users.

3.5.4 Reading level of the clicked results

One of the most noticeable factors in children development and its relation to the web search behavior is an improvement in reading skills. As children improve their reading proficiency they will be able to (i) make sense of a wider range of web results, and (ii) potentially understand better the various elements of a web search engine, such as query suggestions or advertisements.

Collins-Thompson et al. [2011] described a method to personalize web search results based on the readability of the content. They also analyzed the readability of pages accessed by queries targeting content for children, and pages accessed by all type of queries. They found that the former set of queries clearly lead to pages with lower readability complexity. They also found that the higher the readability complexity of the snippets of a web results, the higher is the likelihood of the user to abandon the Website quickly.

To retrace and quantify the improvement in reading level in our data set, we mapped clicked result pages to a 3-scale reading level using Google's "annotate results with reading levels" option.[1] Here, we simply issued the url of the page of interest as a query to Google. In cases where the full url did not return any results, or at least no results with an annotated reading level, we used backtracking by iteratively chopping of parts from the end of the url, hopefully finding a shorter url for which information could be obtained.

Table 3.10 gives a few examples for each of the three reading levels. Note that a single host such as `http://en.wikipedia.org` can host pages of all three reading levels. We calculated averages across all web pages irrespective of the corresponding query volume: 51.6% of the urls were classified as "basic", 35.8% as "intermediate"

[1] `http://www.google.com/support/websearch/bin/answer.py?hl=en&answer=1095407`

reading level	example urls
basic	`www.funbrain.com`, `en.wikipedia.org/wiki/Toy_Story`
intermediate	`en.wikipedia.org/wiki/John_Wooden`, `www.sprint.com/`
advanced	`www.answers.com/topic/mathematics`, `www.merriam-webster.com`

Table 3.10: Examples of websites for each of the three reading levels.

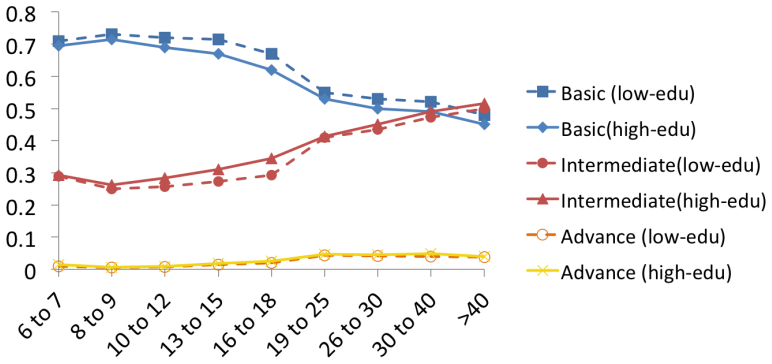


Figure 3.13: Reading level across age and average educational level.

and 2.9% as advanced. For 10.2% we could not obtain a reading level with the current approach.

We observed a general and strong trend for the fraction of clicks on "basic" reading level pages, which declines for older users. At the same time, we observed a weak increase for the "advance" level and a strong increase for the "intermediate" level. We also broke down users according to the education level in their self-reported ZIP code, in order to understand which other factors, apart from age, influence the preferred reading level of users. Concretely, we used the census feature "percentage of population of the age of 25 or higher holding a bachelor degree or higher". We sorted users according to this feature and investigated the lowest 20%-tile, and the highest 20%-tile. Figure 3.13 shows for the fraction of basic reading level pages that children from well educated areas had about 3 years of advantage over children from poorly educated areas. For example, a child from the age range 16 to 18 has a fraction of basic result clicks of 65%. This is slightly lower than the fraction for children in the age range 13 to 15 from well-educated areas, which is 66%, and much higher than the fraction of 60% for other children in the 16 to 18 age range also coming from well-educated areas.

Statistical significance was tested for each proportion (in terms of reading level and educational level) and the proportions were tested in pairs between the ages ranges. All values were statistical significant with p-values < 0.001.

3.5.5 Query and Click vocabulary

We estimated the vocabulary size by counting the number of distinct words employed in a sample of *10K* queries. To make the comparison fair across age ranges we employed a uniform random sample of the same size for each one of the age groups studied. An analogous procedure was carried out to quantify the *web resource* vocabulary. The *web resource* vocabulary is the average number of distinct urls (or domains) that users of a certain age range access. Larger *domain* and *url* vocabularies indicate more diversity of topical interests and greater capabilities for browsing the Internet. A larger number of domains accessed means that a user has explored a bigger portion of the Internet.

The results obtained for both, query and *web resource* vocabulary size, are displayed in Figure 3.14. We found a clear increasing trend in the vocabulary size of users aged 7 to 25 years old. After this age , the vocabulary stops increasing and normalized. Nonetheless, the ratio difference between young children and young adults is less marked than the vocabulary gap reported in natural language. For instance Moore and Bosch [2009] reported that the average number of words known by a 6 years old person is 14K while for 16 years old is 40K. However, this might be explained by the fact that we are measuring the vocabulary of web queries, which are significantly smaller than standard documents. Documents are traditionally employed to estimate vocabulary sizes in natural language.

This result also provides evidence that young users have more difficulties than adults to explore the web given that, on average, they accessed a smaller number of web sites. Note that this is hardly due to a shorter exposure to the Internet since children have been reported to access the Internet from home and even from smartphones. Similarly, the time they spent online has increased dramatically during the last years[1].

This is consistent with our previous findings regarding the high click bias of young users. Recall that we found that young users tend to click more often than adults on top ranked results, which reduce the potential number of websites they explore. The trends found for the vocabulary sizes can also be explained by the fact that young users are interested in a smaller number of topics, as it was shown in the previous sections (Section 3.5.1).

From these results, we can conclude that the portion of the internet that young users explore is significantly smaller than the portion that adult users explore. We believe search engine designers should provide search assistance mechanisms to reduce the query vocabulary gap. Mechanisms to help children find new urls are also needed since they are clicking on a smaller set of urls. For instance, web resource

[1] `http://stakeholders.ofcom.org.uk/binaries/research/media-literacy/oct2012/main.pdf`

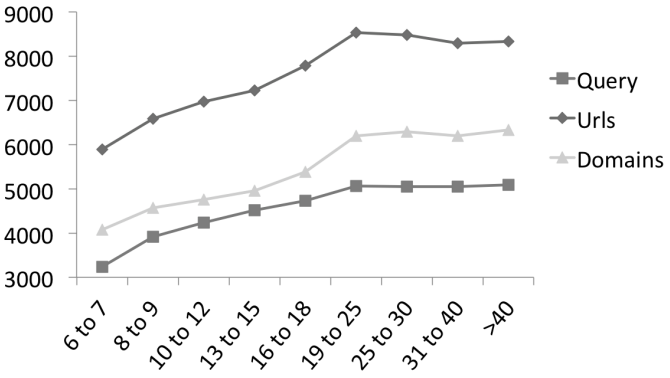


Figure 3.14: Vocabulary size across age groups.

recommendation for young users could potentially aid them to explore more content, and especially it can help them to go beyond the simple web result list, which may not be the best approach to deliver web content for young audiences.

We tested for statistical significance differences between the proportions for all the age ranges pairs. We did not find non-statistical results using the paired t-test (all the p-values were < 0.001). We repeated the process for the query, urls and domains vocabulary size, all with similar results.

3.6 Comparison with AOL search log analysis

In Chapter 2 was carried out an analysis of queries and sessions derived from the AOL search logs and the Dmoz Kids and Teens section. Recall that in the previous chapter we reported micro-averages and macro-averages (in Appendix A), both showing equivalent trends, which allows us to compare the results with the macro-averages reported in this chapter.

When comparing the results from both chapters, it is important to keep in mind that the averages estimated in the previous chapter are representative of users searching for content for children, and particularly of users that clicked on content that is known to be suitable for children, using as gateway the urls of the Dmoz Kids and Teens section. On the other hand, in this chapter the results represent average search behavior of users in a specific age-range. Thus, differences in the trends found for some of the analysis presented in both chapters do not imply that the queries presented in the previous chapter were not submitted by children.

We found different trends for the analysis of query length, click distribution and session length. For the case of query length, we observed in this chapter (Section 3.4.1) that queries are shorter for the case of young users (both children and teenagers) while in the previous chapter (Section 2.6) we observed longer query length averages

for the case of the queries identified through the *kids* and *teens* Dmoz urls. These results indicate that longer queries (in respect to the average web user) are required to find moderated results for young users while on average young users submit shorter queries, as it was shown in this chapter. This result provides evidence that young users have difficulties accessing the right content to satisfy their information needs.

Similar observations can be applied to the differences found in the click distribution. In Section 2.6.7 we observed a higher amount of clicks on lower ranked results from users searching for content focused on young users, while in Section 3.4.3 we observed the opposite trend, greater proportion of higher ranked clicks on young users. This result also points to the difficulty of finding appropriate content for young users since high quality content for this audience was often located at lower ranked positions. This situation is also reflected in the session length results found in Section 2.7.1 and 3.4.8, in which we observed that longer sessions (greater amount of clicks) are required to find and consume high quality content for young users while these users have shorter search sessions than the average web user, which also points to search frustration.

On the other hand, we found similar trends in terms of vocabulary size, natural language usage and topic distribution. In sections 2.6.2 and 3.4.2 was observed that the vocabulary size of users aged 8-9 and 10-12 were 3.9K and 3.2K respectively, while the vocabulary size observed in the previous chapter was 3.7K for the *kids* queries. The trends were consistent in both analysis in the sense that the vocabulary size of teenagers and adults were larger than children. Nonetheless, we found bigger vocabulary sizes for teenagers and adults through the AOL logs: 7.7K against 4.5K for the case of teenagers, and 11K against 5K for the case of adults (or average web user). In sections 2.6.2 and 3.4.2 were found larger proportions of question queries in the set identified through the Dmoz urls (and in the queries submitted by young users), than in the queries submitted the average web user. Similarly, the usage of natural language in the queries was larger for the former groups of users.

3.6.1 Topic distribution comparison

In this section we quantify the correlation between the topic distribution of the queries identified in the previous chapter and the queries of users submitted by users of the Yahoo! logs. The motivation of this analysis is to determine if the queries identified through the Dmoz urls are representative, from a topical point of view, of the queries submitted by children and teenagers. Recall that the vocabulary size trends observed in both datasets (AOL and Yahoo! Search logs) were consistent, which provides evidence to this claim.

We followed an approach analogous to the one presented in Section 3.5.1. We estimated the Pearson's correlation coefficient between the topic distribution obtained

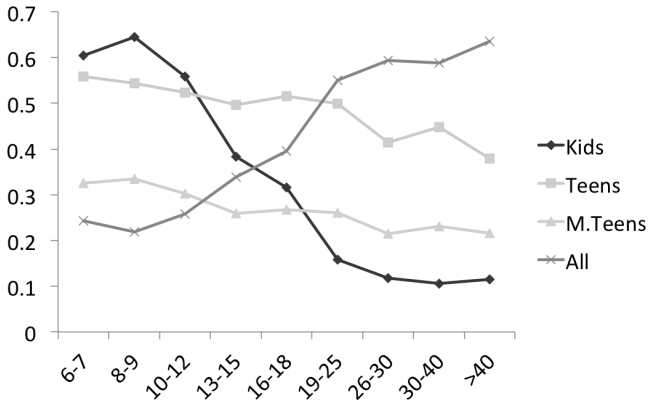


Figure 3.15: Pearson's correlation between the topic distribution of the queries identified through Dmoz and the topic distribution of the queries of users of different ages.

with the queries identified in the previous chapter and the topic distribution of the queries submitted by each one of the ages ranges. Recall that both distributions are obtained by mapping each query to one of the categories of the Yahoo! Directory, thus the topics in the distributions are consistent. Figure 3.15 presents the results obtained. For the case of the queries identified with the *kids* urls we found, in general, that the younger the user, the higher the correlation coefficient. The topics represented by these queries correlate the best with users aged 6-7, 8-9 and 10-12 for which we obtained substantial agreement. The highest correlation is obtained with the queries submitted by users aged 8-9 years old (0.65). We observed that the topics of these queries correlate moderately with the queries submitted by teenagers and have low correlation with queries submitted by users over 19 years old.

A clear trend was also observed for the case of the distribution obtained from a sample of all the queries in the AOL, which represent the topic of the average web user. In this case the older the user the higher the correlation. Substantial agreement is obtained for users over 26 years old. The highest correlation is obtained for users over 40 years old (0.64).

For the case of the *teenager* queries the trend was not clear. Even though the queries identified through this type of urls had the highest correlation with the queries submitted by teenagers aged 13-15 and 16-18, we also observed strong correlation with the queries submitted by users under 15 years old. The correlation decreases for users over 26 years old. Lower correlation values were observed for the topics mapped from the queries obtained with the *Mature Teens* urls of Dmoz.

On overall, we found that the highest correlated data sets are aligned with the desired age range. That is, the *kids* queries fit better the topic of the queries of users aged 6-7, 8-9 and 10-12. Similarly the *teenager* queries correlate better with the

queries of users aged 13-15 and 16-18. The exception to this trend was the queries obtained with the *Mature teens* urls. This result may be due to the fact that the Dmoz section focus on the youngest population segment (*i.e. children).*

This result provides evidence of the adequacy of using the queries identified through the *Dmoz* urls to represent the topics targeted by topics of young users. The similar trends observed in natural language usage and query size vocabulary also support this statement. Note, that the queries extracted from the AOL log are also representative of successful queries. We will extract query reformulations from both datasets (AOL and Yahoo! Search logs) to address the problem of query recommendation in the next chapter. We will show consistent results when using both datasets.

3.7 Conclusions and future work

3.7.1 Findings summary

In respect to research question *RQ-2.1*, it was observed clear evidence of search difficulty in young users. The shorter average query length observed along with the greater usage of natural language shows that these users have difficulties formulating specific queries with keywords, which are the main mechanism to query in state-of-the-art search engines. The larger proportion of short clicks also indicates search difficulty, since it shows that pages are abandoned quickly. In terms of click behavior the position click bias for children is worth pointing out.

This bias also leads to a higher fraction of ad clicks and to a higher fraction of cases where (useful and correct) spelling corrections are undone by the user when clicking on, say, "Show only '*brittnay spears*' ". This result shows that young users have a tendency to "click whatever is presented at a prominent position" which has implications for the design of appropriate search interfaces. This behavior, along with the trust young users have on the veracity and quality on the content published on the Internet, contributes to the clicks on content that is not suitable for young users. At it as observed with the greater proportion of accidental clicks on adult content, and the greater large amount of clicks on advertisement, which potentially lead to more search frustration.

In respect to *RQ-2.2*, we found notable difference between children and adults in terms of search and browsing behavior. However, it was surprising to find small differences between children and teenagers. For instance, users between 16 and 18 behaved more like children between 8 to 9 than young adults in the 19 to 25 age range. This "sudden jump to adulthood", although not for all features, could potentially be explained by children leaving home and starting college or a job. Some features in which children and teenagers differ by a large margin were: topic distribution,

query vocabulary size and the proportion of question queries (with natural language) submitted to the search engine.

For *RQ-2.3* we observed that several aspects of children and teenager development were reflected in the search logs. For instance, we found a direct correlation between the topic distribution and age. Interestingly, we observed a wide variety of topics in the distribution of adults. On the other hand, for the case of children and teenagers the queries concentrated on a small set of topics. Stereotypical topics such as gaming and entertainment were observed in children and teenagers respectively. However, we observed an unexpected higher interest in emailing and other Yahoo! products such as Yahoo! Answers for the case of children. The latter was also prominent for the teenager segment of users.

The development of children was also reflected in clear differences in the query vocabulary size and the readability level of the content accessed. Features that we hypothesized could reflect human development, such as the sentiment express in queries, were not conclusive and more research is required in this regard and other effectual aspects of search.

3.7.2 Recommendation for the development of IR technology for children

A call for better query assistance functionality has been pointed out in previous research [Bilal, 2002; Druin et al., 2009]. In our work, we also found evidence of the need of this functionality. From our observations, young users (particularly under 12 years old) would greatly benefit for query expansion and query recommendation mechanisms, since their average query length is the smallest from all the age groups. They also had the smallest query vocabulary size. Query suggestions with information aspects (i.e. query senses) tailored with topics of interest for children would greatly help children to narrow down their searches, and it would improve their chances to find information that is on topic. These tools would also improve the chances of returning results that are better suited for them. We expect that high quality suggestions would greatly increase the usage of these tools by children.

For the case of teenagers we found interesting the large usage of question queries, which suggests the need of providing robust mechanism to parse this type of queries in order to understand in deep their information needs. Alternatively, resource selection mechanisms to identify when it is appropriate to return results from specialized question answering systems based on the topic of the query are also relevant, especially since we observed that children and teenagers make extensive use of these services (e.g. Yahoo! Answers).

The main topics identified in the topic distribution for each age range are valuable for the design of aggregated search interfaces, since the more prominent topics can

be associated to specific verticals (e.g. games, school, entertainment). In this regard, resource selection methods and novel aggregation paradigms for young users are worth to be explored.

3.7.3 Recommendations for future research

Even though we collected clear evidence to address the research questions posed in this chapter, several open questions still require further research. We characterize the search behavior of users based on their age, however it is unclear how this behavior varies based on the topic that is searched and the specific type of information need (e.g. open and close informational needs). A natural follow up of this work would include a search analysis conditioned not only on age but also on topic.

We explored the proportion of clicks on content of different reading difficulty by considering the age and the average educational level in which the user is located, by combining the search logs and the U.S census data from 2000. The exploration of the search difficulties and search behavior differences of users within other demographic features is a highly relevant research line that can provide clues on how to improve the search experience of specific niche of users. For instance, knowing the specific search difficulties of users per educational level, region and income would lead to the possibility of displaying, if the user wants to, specialized results that address specific search difficulties. Note that we are not suggesting to blindly show users specific results according their demographics, instead we suggest to provide the option to retrieve these specialized results to all the public.

Demographic features such as race and language proficiency require special attention to provide adequate tools aimed at improving the content readability. Cultural differences in terms of search behavior, and particularly cultural topic preferences represent another important, yet not well-studied, area in information retrieval.

The impact of aesthetic characteristics of the websites visited by young users is yet to be explored, particularly on the large scale and its relation to the problems that arises during the search process of young users. For instance, page layout, amount of multimedia content, number of images, font types, font sizes, text and background color are all features that need to be explored in future studies. Measuring the impact of these design parameters would lead to a better understanding in how to design content that is more engaging for children.

More research is also needed to understand the affective factors that arise during the search process for the case of very young users. We quantify the amount of sentiment in the queries, however no trends were found. Quantifying the sentiment in the content accessed by children and exploring its correlation to the click duration can provide a better understanding in this respect.

As it was pointed out in the previous section, research towards mechanisms to

automatically select search services relevant to the topic based on the query and age of the user would be highly beneficial. New paradigms to aggregate content from these services are also needed to simplify the exploration of results and the engagement during the search process.

Chapter 4

Browsing behavior of Young Users: Search triggers

This chapter is based on Duarte Torres et al. [2014a].

4.1 Introduction

In this chapter (as a follow up to the previous chapter) is characterized the type of activities that young users carried out online in the Internet when a search engine is not being used. We also analyze which of these activities are more likely to motivate search. The understanding of the online activities engaged by young users and how they relate to their search queries lead to an integral picture of search engine usage. At the best of our knowledge no research has been carried out in this regard that is both: on a large scale and unobtrusive. It is important to mention that recent research [Cheng et al., 2010; Kumar and Tomkins, 2010] has addressed the characterizing of the browsing activities engaged by the average web user, however these studies are not oriented towards the young population and in general the focus is not on the age demographic dimension of the users.

4.1.1 Research questions

The contribution of this chapter can be summarized by the following research questions, which address the research aims expressed previously:

- **R.Q-3.1**: What activities are carried out by young users on the web browser besides web searches?, How prominent is browsing for each age range? At what ages are multimedia searches preferred?
- **R.Q-3.2**: Which types of search and browsing activities are more likely to trigger searching the web and multimedia search engines in the case of young

users? Do these triggers differ from those observed in adults users?

As it was the case with research questions *RQ-2.1* to *RQ-2.3*, research questions *RQ-3.1* and *RQ-3.2* are addressed by analysing a set of carefully chosen log metrics estimated on a per-user basis. However, these two research questions are addressed using a large sample from the Yahoo! Toolbar logs instead of the Yahoo! Search logs. The toolbar logs contain records of the urls accessed in a browser, which may include search activity and browsing activity (e.g. emailing, social networking websites). All the toolbar data is captured only after the user has explicitly opt-in to have their browsing activity registered. All the data employed in our studies has been previously anonymized. This data allow us to analyze the online activity of users of different ages outside the search engine, including activity before engaging web search. The log data employed correspond to the same market and time window of the query logs used to address *RQ-2.1* to *RQ-2.3*.

In this thesis, we refer to browsing activity as all the action that is not carried out within the scope of the search engine. Thus, with *browsing* we refer to website visits such as email portals, social networking sites (e.g. Facebook), news portals (e.g. BBC), head-listing websites (e.g. Ebay) and in general any website that is accessed directly by typing the url in the browser or through a bookmark. Note that this differs from the *search by browsing* definition addressed in previous case-studies of children search behavior [Druin et al., 2009], in which children are tested to engage informational tasks under two settings: keyword search (i.e. search engines) and by browsing a hierarchy of categories. For us *browsing activity* refers to browsing the Internet in general and not necessarily to search information through a hierarchy of categories.

To address *RQ-3.1* we estimate the volume of search and browsing activity for all the age ranges. We analyze the ratio between search and browsing activity to contrast how users of different ages spend their time in the Internet. A more detailed understanding of the browsing activities is obtained by estimating the likelihood of a user to submit a query in a search engine given that their previous activity was a browsing activity. Particularly we analyze web search and *multimedia* search separately. An important feature of the toolbar logs is that the interaction on multimedia verticals such as Yahoo! Images and Yahoo! Videos are also registered. We hypothesize that young users carried out a greater amount of browsing activities than older users, since the latter have a greater expertise with search engines.We also hypothesize that multimedia search is preferred by young audiences since the rich content found in these web services is more engaging.

RQ-3.2 is addressed by analysing the likelihood of web and multimedia search to trigger other searches. A more detailed analysis of the likelihood of carrying out search after a browsing event is carried out to address this research question. Concretely, we brake down the cases in which the search query is explicitly mentioned in the

browsing website, the cases in which it is mentioned in its domain, and the cases in which it is not mentioned at all. This analysis provides direct evidence of the specific types of browsing activity that lead to query submissions in a search engine.

We hypothesize that search in each age range is motivated by specific type of browsing events. For instance, we expect knowledge intense websites (e.g. Wikipedia) to be more likely to trigger searches of young users (e.g. children) than of adult users. Similarly, we expect social networking page views to trigger search more frequently in teenagers.

4.1.2 Chapter Organization

Related work is presented in Section 4.2. Section 4.3 describes the Yahoo! toolbar sample employed in this study and the data cleaning steps. A description of the method to extract and analyze measures from this data is also addressed. Particularly, we define the type of events that are considered for this part of the study. Section 4.4 presents the discussion regarding the session characteristics found in the toolbar logs. These characteristics are contrasted along the age demographic dimension (this refers to *RQ-3.1*). Section 4.5 discusses the likelihoods of different browsing page-view and the search activity that lead to web and multimedia search. In Section 4.6, we expand the analysis carried out in Section 4.5 by quantifying the cases in which the query is mentioned in the url and domain previously browsed. Sections 4.5 and 4.6 addressed *RQ-3.1* and *RQ-3.2*. Finally, Section 4.7 concludes this chapter with a discussion of our main findings and how they could be applied to state of the art search engines. We also provide recommendations for future work.

4.1.3 Limitations of this study

Analogous to the limitations pointed out in the previous chapter, detailed comparisons of browsing behavior within adult age groups were left out since the focus of this thesis is on the young user segment. As it was the case with the search logs, the results presented in this study are derived from log data from the U.S. market, in which the predominant language is English.

4.2 Related work

Kumar and Tomkins [2010] carried out a large-scale user behavior study on search and toolbar data using logs from the Yahoo! system. They studied search sessions of adult users, in general the age dimension was not addressed in their study. They proposed a classification of browsing page views based on content (e.g. news games, portals), communication (e.g. email, social networking) and search (e.g. web, multimedia).

They found that around half of the page views belong to the content categories and a third to the communication category. The chain of referrals (i.e. the chain of pages visited in the Internet browser) was analyzed to characterize the way users navigate through the page views. They showed that mail, news and social bookmarking page views occur in isolation and that on overall 35% did not have any referral. In this chapter, we employed a similar classification of browsing page views. Our work differs from Kumar and Tomkins [2010] in that we analyze the browsing behavior of users not only based on page view content classification but also according the age of the user.

Cheng et al. [2010] presented a method to predict the search intent of users by considering their browsing behavior. They first carried out a user study to characterize the main types of page views that lead to web searches by using toolbar data. Then, they employed a machine learning approach to rank queries according the previous browsing activities of the user. In this chapter we are interested in understanding which type of browsing activities lead to web and multimedia search and how these activities differ with the age of the user.

Goel et al. [2012] carried out a large scale study of browsing behavior of users of different demographic characteristics by employing web panel data and user-level demographics such as age, sex, race, education and income. In terms of age their study included users from 7 years old to 80 years old. They found that all demographic groups spend the majority of their time online on the same type of activities: social media and emailing. In this regard, they observed that teenagers aged 15 use the most social media and its usage decreases consistently after this age. On the other hand, they observed that emailing usage correlates positively with age. They also found pronounced differences regarding the frequency in which different demographic groups (in terms of salary income and education) access websites with news, health and reference (encyclopedia) content. They also explored the digital divide in terms of Internet usage for research and informational queries, its relation to the educational background of the users and other demographic features such as salary income, race and gender. Our work differs from theirs in that we focus on identifying the browsing activities of very young users and how they differ from adults. We also emphasize our analysis in the browsing activities that motive search.

On overall, the previous research on browsing behavior in the Internet has focused on providing a framework to classify the possible page views, and to provide automatic methods to predict a page view type given the previous pages accessed by the user [Cheng et al., 2010; Kumar and Tomkins, 2010]. A characterization of the frequency of each type of page view has also been explored for the case of adult users [Cheng et al., 2010; Kumar and Tomkins, 2010]. Nonetheless, none of these studies focus on young users, leaving several research questions open for comparing the browsing behavior of children, teenagers and adults.

4.3 Method

In this section is described the Yahoo! toolbar data employed to address research questions *RQ-3.1* and *RQ-3.2*. To address these research questions we followed the TLA (Transaction Log Analysis) framework, as it was the case with the Yahoo! Search logs. Recall that this framework is based on the collection, preparation and analysis of log data. Each one of the items is described in the following paragraphs.

4.3.1 Toolbar data collection and preparation

We employed logs from the Yahoo! toolbar. The toolbar is a browser application to aid users search and browse the Internet without accessing directly a search engine. The main advantage of analysing the search behavior using logs from the toolbar is that the activity outside the search engine is captured. Traditional search logs only capture the queries and clicks registered during search sessions, on the contrary, the toolbar logs captures searches across different search engines (not exclusively a search engine product) and additional browsing activity carried out before, during and after the search activity.

We collected Yahoo! toolbar log entries for each age range over a time window of four months during 2010. Note that we employed the same time window employed for the Yahoo! Search logs and the same age ranges specified in Section 3.3.1 of the previous chapter. We collected in the order of tens of thousands of entries for users up to 12 years old, and hundreds of thousands of entries for users over 13 years old. It is important to mention that exact numbers are not reported since they are considered business sensitive information. For the preparation of the data we only employed log entries from users that agreed explicitly to be logged according the usage conditions of the toolbar application[1], and only from users that have a valid Yahoo! account. We employed the definition of valid account described in the previous chapter (Section 3.3.1). Similarly, queries that could reveal personal information (such as non-frequent names, credit card numbers, telephone numbers) were anonymized prior to the analysis and discarded. It is important to mention that very large volumes of data originates in the Yahoo! Toolbar from logged in users, thus the results presented in this chapter are still representative of the underlying population.

Each entry in the toolbar logs has the following structure: {*user ID, timestamp, url, referrer url*}. The *user ID* and *timestamp* were employed to sessionize the data as we did with the search logs. The *referrer url* is the previously web resource accessed by the user and the resource from which the current *url* was found in the entry log. The *referrer url* is used to track consecutive stream of events in the toolbar logs.

[1] http://info.yahoo.com/legal/us/yahoo/toolbar/tbeula/tbeula-282.html

Contrary to the standard search logs, tracing a linear set of events is not trivial on the toolbar logs since the user can have several browsing tabs open, and may carry unrelated search activity on each one, which are registered in the logs as interleaved events [Huang et al., 2012].

4.3.2 Yahoo toolbar log analysis

For all the analysis, we created user sessions using a time window of 30 minutes between two consecutive log events. The same strategy has been employed in previous browsing web behavior studies [Cheng et al., 2010; Kumar and Tomkins, 2010]. All results presented from the toolbar logs are based on aggregated statistics to preserve the anonymity of the users.

In respect to *R.Q-3.1* we hypothesize that the session characteristics in the toolbar differ by age. Particularly that the amount of search engine usage and Internet browsing changes according the age range of the user. For this purpose, we analyze the sessions of users in the toolbar logs and their characteristics. Specifically, we looked at the length and duration of the sessions and at the proportion of search and browsing activity that is carried out within sessions.

For *R.Q-3.2* we analyze the toolbar sessions in two dimensions: by type of browsing activity and by type of search activity. For the former we employed the taxonomy suggested by Kumar and Tomkins [2010] to classify the possible browsing events that can occur before search activity. For the latter we make the distinction between web and multimedia search. The latter refers to queries submitted to the Yahoo! Videos or Yahoo! Images search services. Switch patterns of browsing activity to search activity were quantified using these classifications. These patterns tell us which activities are more likely to occur before search events.

In the following sections we describe in detail the methodology and results obtained for each one of the analysis mentioned above. We will refer throughout these sections to the following definitions:

Definition 1. ***Event****: A event refers to a user entering an url in the browser, clicking on a link or entering a query into a search engine. Concretely is represented as a tuple {user ID, timestamp, url, referrer url}, which is also the representation of an entry in the toolbar log. It is important to mention that redirect links and other log artefacts are not considered events (these entries were properly removed from the data collected).*

Definition 2. ***(Search) query event/ (Search) Portal event****: A search event occurs when the url matches any of the major search engines (google, yahoo, bing, aol, ask) and contains a search query (e.g. www.google.com/search?q=elmo). A search portal event occurs when the url match any of the major search engines but no search query is detected in the url (e.g. www.google.com)*

Age	Events	Search e.		Browsing e.		Portal e.	
		Volume	Perc.	Volume	Perc.	Volume	Perc.
6-7	31.24	4.67	14.95%	23.55	75.38%	3.02	9.67%
8-9	38.64	4.82	12.48%	30.07	77.82%	3.75	9.70%
10-12	39.36	4.42	11.23%	31.35	79.64%	3.59	9.12%
13-15	55.19	5.29	9.58%	45.78	82.94%	4.12	7.47%
16-19	59.51	5.38	9.04%	50.09	84.18%	4.03	6.78%
19-25	60.16	8.73	14.52%	47.09	78.28%	4.33	7.20%
26-30	56.50	8.28	14.66%	44.35	78.49%	3.87	6.85%
31-40	49.38	7.73	15.65%	37.53	76.01%	4.12	8.34%
>40	41.89	7.29	17.41%	29.89	71.36%	4.71	11.24%

Table 4.1: Toolbar session and avg. number of search, browsing and portal events per user session and their percentages in respect to the total number of entries.

Definition 3. ***(Search) result event****: A search result event corresponds to any url clicked from the list of web results obtained for some search query. We detected these events using the referrer url. In other words we classify the event as a search result event if there exists at least one chain of events connecting the target url to a search query event.*

Definition 4. ***Browsing event****: A browsing event is any url that is not a search query, search Portal, or search result event. We also disregard urls containing queries even if the urls do not match any of the major search engines (e.g. www.facebook.com, www.bbc.com).*

It is important to mention that the results presented in this section were proven to be statistical significant when comparing each one of the children and teenager age ranges against the group of adults using the two-tailed t-test at a 0.1% level. As it was the case with the search log analysis, we report p-values in each section for values that were not proven statistical significant. However, given the size of the data most of the results were statistical significant with very small p-values (< 0.001)

4.4 Session usage and characteristics

The macro-averaged duration in minutes of the sessions and their sizes in terms of number of events were estimated to provide an overview of the browsing behavior of young users in the internet. We also estimate the proportion of search events against browsing events across the age ranges.

Table 4.1 presents the results. We found that the sessions are longer for older users, particularly for users between 19 and 25 years old. On average the sessions of users

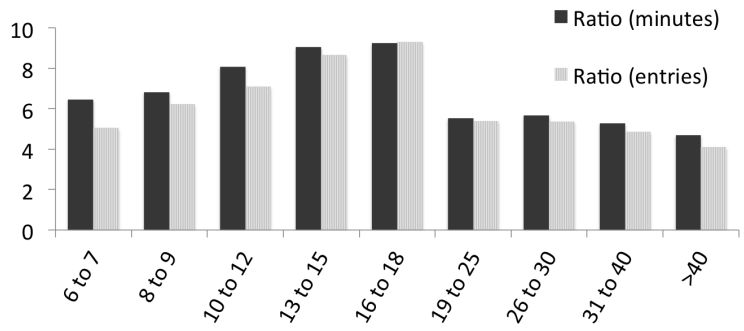


Figure 4.1: Ratio of browsing activity against search events in terms of number of events and duration in minutes.

from 7 to 9 years old contain half the number of events than the average session length of users from 19 to 25 years old. The size of the teenagers' sessions were comparable to the session size of adult users.

The larger sessions of adults is reflected in both greater search and browsing activity. However, we observed that adults have a bigger proportion of search events in respect to browsing events. For instance, for the case of the sessions of users aged 10 to 12, 11% of the events are search events and 80% are browsing events, while for users over 40 years old the proportion of search events is around 17.5% and 71% of browsing events.

Interestingly we also observed a bigger proportion of search activity for the case of users between 8 to 9 and 10 to 12. For teenagers (users between 13 to 15 and 16 to 18), the gap between search and browsing activity is higher: 9% of search events against 84% of browsing events. For the youngest group of users we did not observe large differences in the proportion of search and non-search activity in respect to the sessions of adult users.

Kumar and Tomkins [2010] also explored the proportion of browsing and search activity. Similarly, they reported significant larger amounts of browsing activity in respect to search and portal activity, as we found for most of the age ranges. They reported that on average 9% of the page views correspond to search activity and 21.4% of the page views are derived from search events directly (query submission) or indirectly (browsing a web result). Averages with absolute number of events are not reported. In our findings, the average proportion of search events for users aged 10 to 12, 13 to 15 and over 40 years old were 11.2%, 9.5% and 17.4% respectively.

In general we observed that teenagers are the group of users with the smallest proportion of search, while children and adults had the biggest proportion of search in their sessions. This result is interesting because we were expecting a higher percentage of search activity for the case of teenagers, since these users have greater search capabilities than children.

The lower proportion of search activity found in the children and teenager sessions can be explained by the fact that these users have a strong interest in a few number of topics, as it was shown in the previous chapter in Section 3.5.1. For instance, gaming is a predominant topic for users under 12 years old. We believe that specialize websites dedicated to this topic can be accessed directly by typing the url in the browser or by bookmarking the website. Recurrent gaming websites such as *Club Penguin, PopTropica and Nick Jr.* are frequently accessed websites by children [1] and it is reasonable to expect that children access these websites directly without searching for them in a search engine.

From the topic preferences observed for teenagers, we believe that a strong bias towards social networking sites (e.g. Facebook, MySpace) lead to the highest proportion of browsing activity, since these web services are easily accessible without the aid of a search engine.

Figure 4.1 depicts the ratio of browsing activity against search activity in terms of session duration in minutes and session length (number of events). We observed that users from 10 to 12, 13 to 15 and 16 to 18 years old carried out 8 times more browsing activities than search activities. For the case of grown ups, it was found that users spent 6 times more browsing than searching. For users under 10 years old the ratio was around 5. Interestingly, the session duration ratios were more accentuated than the ratios of the session lengths, which shows that children spent more time exploring the urls.

We also explored differences in the search activity of users by looking at the average proportion of web results and queries submitted within the sessions. We found that a larger number of web results are explored by adults, which is consistent with the findings reported in the previous chapter. For instance, the average number of unique queries per session found for users aged 10 to 12 was 0.8, while the number of unique search results (unrepeated search results independently of the query) was 1.5. For users over 19 years old the average found were 1.1 and 3.4, respectively. Note that some values can be lower than 1 because we are accounting for unique queries submitted within the search session and some sessions can contain only instances of the same query or no queries at all. As it was mentioned before, this result indicates that young users explore less the list of web results. This result can also indicates less satisfactory searches for the case of young users since less search results are clicked, given that a similar number of unique queries is submitted within the sessions across all the age ranges.

We tested for statistical significance all the macro-averages reported in Table 4.1 (number of search events, browsing events and portal events). All the paired tests were found statistical significance with very small p-values (< 0.001). This was also the case for the average duration in minutes, depicted in ratios in Figure 4.1. The

[1] `http://www.ebizmba.com/articles/kids-websites`

only value that was not found statistical significant was the average duration for the pair 13-15 and 16-18 (p-value of 0.064)

4.5 Event to search query switch patterns

We extracted from the toolbar sessions all the pairs *event* → *search query* in order to understand the events that are likely to occur before search events. *Event* may refer to a *query event*, *browsing event* or the *start of the session*.

For the *query events* we make the distinction between queries submitted to the standard web search vertical, and queries submitted to the image and video verticals. We will refer to the queries submitted to these two verticals as *multimedia queries*. The motivation for making this distinction is that children have been shown to prefer visual search for certain search tasks [Druin et al., 2010]. We hypothesize that this phenomenon is also reflected in the usage of verticals with visual content.

Cheng et al. [2010] proposed an automatic method to predict search intent based on user browsing activity. They found from the Yahoo! toolbar logs that on average (no age differences were accounted) 19% of the toolbar sessions contained browsing to query search patterns, while 24.3% contains search activity and 75.7% of the sessions contained only browsing activity. We will show that these values are consistent with our findings for the case of adult users, although we observed smaller percentages for the case of young users.

We extended the switch patterns defined by Cheng et al. [2010] to include the usage of multimedia search queries. The definitions are described as follows:

Definition 5. *Start of the session → (web/multimedia) query: This pattern occurs when the search query is the first event of the session.*

Definition 6. *Web/multimedia query → (web/multimedia) query: This pattern occurs when the event before the search query is a search query. For the analysis, we only consider the cases in which the previous query is different to the current search query.*

Definition 7. *Web result event → (web/multimedia) query: This pattern occurs when a search result event leads to a new search query event. Note that the query of the new search event needs to be different that the query utilized to retrieve the current web result event. We include this pattern because it is reasonable to expect a significant amount of cases in which web results trigger new web searches.*

Definition 8. *Browsing event → (web/multimedia) query: This pattern occurs when a browsing event lead to a query search event.*

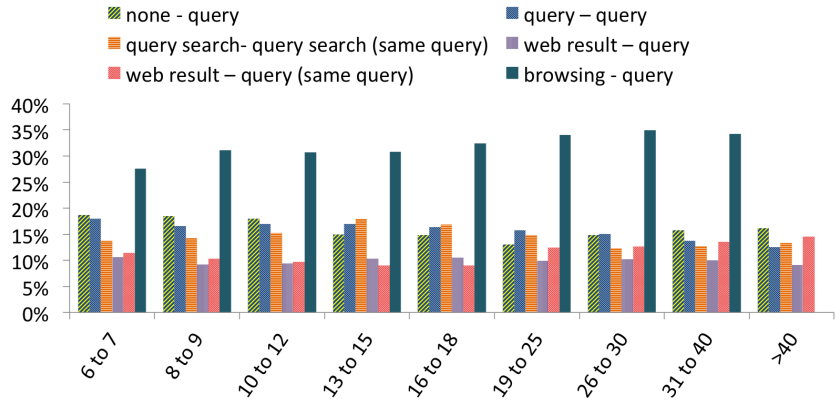


Figure 4.2: Proportion of search patterns for all the age ranges.

We followed the taxonomy of page views employed in [Kumar and Tomkins, 2010] to distinguish between the most relevant types of browsing events. Concretely we defined the following browsing event types:

1. Head listings: events in which the *url* domain is either Amazon, Ebay or Craiglist.
2. Mailing: events in which the *url* match any of the major email service providers.
3. Social: View on websites on any of the major social networking websites. We created this list manually by choosing the social networking websites with the greatest traffic volume[1].
4. Knowledge pages: Websites matching the domain *Wikipedia*.
5. Multimedia: Websites from Youtube, Hulu, Flicker, Phototobucket,
6. Other: Websites that could not be classified in any of the previous categories.

4.5.1 Web search triggers

Figure 4.2 shows the proportion of the patterns defined in Section 4.5 for the queries submitted to the web vertical. The proportions reported are normalized over the total number of pairs *event* $\rightarrow$ *web query* for each age group. We found that a large percentage of browsing activity lead to search events in all the age groups, although this proportion is higher for adults (30% for users under 19, and 34% for users over 25). We also observed that a large proportion of search queries were submitted right at the beginning of the sessions (18.7% for users under 12, and 14% for users over

[1] `http://en.wikipedia.org/wiki/List_of_social_networking_websites`

25). Interestingly, the percentage is higher for children, which suggests that these users start more frequently exploring the internet by using a search engine instead of browsing other resources. This may be due to the fact that the web search engine portal is the default website for most of the commercial browsers. This result also shows that search engines play a bigger role for very young users, since online activities start from the search engine more often than for adult users. We also observed that around 10% of the query events were triggered by exploring web results and 15% of the query events were submitted right after the submission of a different query (e.g. query reformulations).

The following patterns from Figure 4.2 were not found statistical significant (at 0.1%) when using the paired t-test between age groups: none-query between 6-7/8-9 (p-value 0.537) and 6-7/10-12 (p-value 0.488); browsing-query between 10-12/13-15 (p-value 0.191) and query-query between 10-12/13-15 (p-value 0.033); web result-query (same query) between 6-7/8-9 (p-value 0.412), 10-12/13-15 (p-value 0.288) and 13-15/16-19 (p-value 0.165). All other result pairs were statistical significant at 0.1% with the paired t-test.

Figure 4.3 presents the likelihood of having a *web query event* given each one of the non-browsing event types. Values reported in Figure 4.3 were estimated using the fraction between the number of pair events $e \rightarrow$ *web search query* over the total number of events e found in the sessions for each age range. For instance, the likelihood of having a *web query event* given a *browsing event* is estimated by dividing the number of pairs *browsing event* $\rightarrow$ *web search query* over the total number of *browsing* events. Consistent with the results reported in Figure 4.2, children had a higher likelihood than adults of starting a session with web search. The likelihood for children under 12 years old was around 0.55. For teenagers the likelihood was 0.45 and for adults 0.42 (this result was obtained by normalizing over the total number of sessions per user). The probability of web search being spanned by a click on a web result is around 0.04 for users over 25 and around 0.06 for users between 10 to 19 years old. These values indicate that *older* children and teenage users need to refine their searches more frequently than older users since the clicks on web results lead to new searches.

Figure 4.4 depicts the likelihood found for the browsing patterns. We found that for most of the types of *browsing events*, the probability of leading to a web search was relatively low for all the age groups. For instance, the probability of having a search event after a page view on a browsing mail event was around 0.08 for children and adults, while for teenagers the likelihood was around 0.05. For the *social* and *multimedia* browsing events the probability was 0.02 for children and adults, and 0.03 for teenagers. Interestingly, this was not the case for the *knowledge category*, which had a significantly higher chance to occur prior to a web search event. The likelihood was around 0.11 for most of the age groups. For the age group 6-7

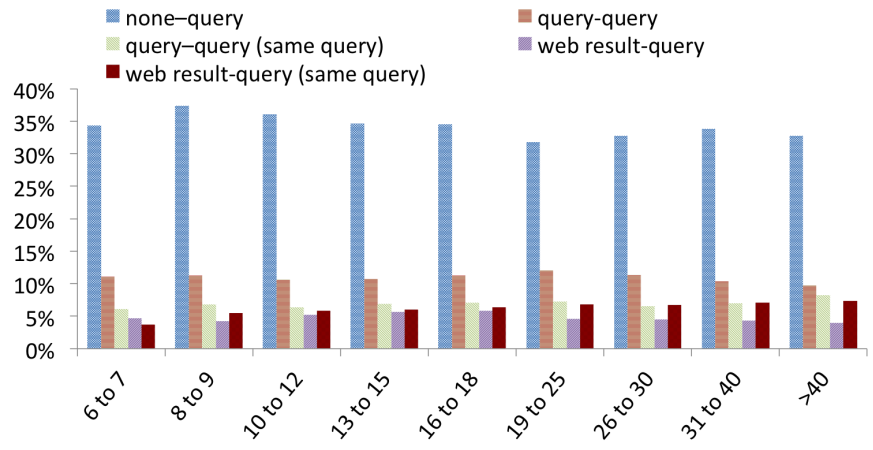


Figure 4.3: Search patterns likelihoods (Web search).

the likelihood was 0.16. This result reflects that very young users are keener to carry out web searches after exploring educational content (e.g. Wikipedia) to satisfy their information needs. Even though this result came as a surprise, other studies have reported that Wikipedia is on the top 12 websites more frequently accessed by children aged 5 to 6[1]. These values may also indicate that this group of users requires more explanations or background information to understand the content available in educational resources. In terms of statistical significance, for Figure 4.3 we found that the pairs 6-7/8-9 and 10-12/13-15 were not statistical significant for the pattern *search query-search query* (p-values 0.143, 0.053) and 10-12/13-15 for the search pattern *web result-same search query* (p-value 0.070).

For the browsing patterns displayed in Figure 4.4, we found that the pairs 8-9/10-12 were not statistical significant for the *knowledge-search query* pattern (p-value 0.266) and the patterns involving multimedia and other browsing for the pairs 13-15/16-19 (p-values 0.113 and 0.094).

On overall, these findings suggest that providing media and content related to the topics of the *knowledge* websites category can be highly beneficial for very young user, particularly for improving the readability of educational content. We believe that this aid can be provided in three ways: *(i)* providing definitions to complex words as it has been suggested in domain specific IR [Azzopardi et al., 2012a]; *(ii)* text simplification as it is suggested by Vettori and Mich [2011], and *(iii)* providing related results from different resources to ease the interpretation of the information displayed by using, for instance, results from different genres that may be more familiar and engaging for children (e.g. images, sounds, games).

[1] `http://stakeholders.ofcom.org.uk/binaries/research/media-literacy/oct2012/Annex_3.pdf`

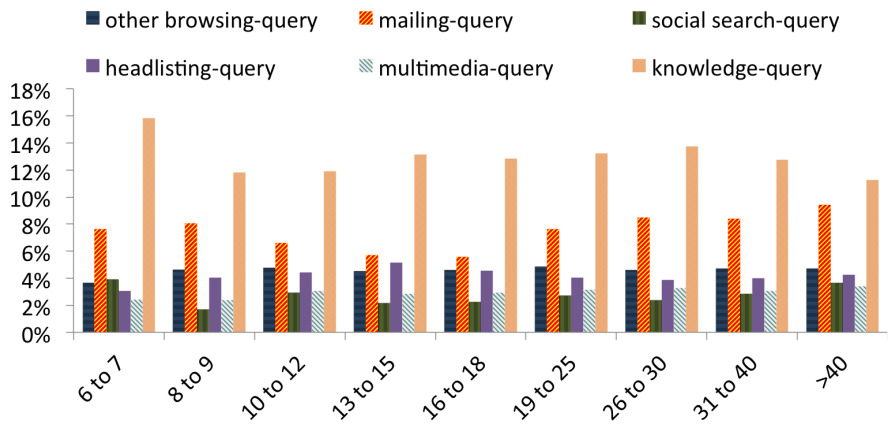


Figure 4.4: Browsing pairs likelihoods (Web search).

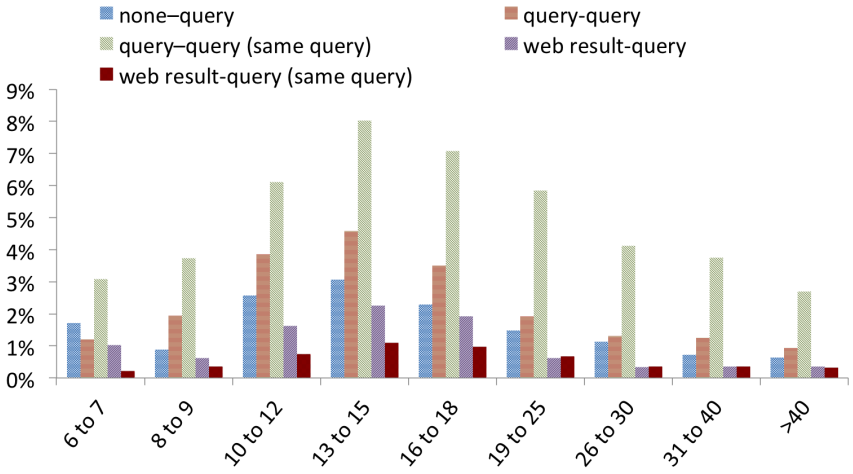


Figure 4.5: Search patterns likelihoods (Multimedia search).

4.5.2 Multimedia search triggers

Figure 4.5 presents the likelihood of having a search event on a multimedia vertical given the non-browsing event types.

On overall, we observed that users between 10 and 19 were 2.4 times more likely to submit queries on a multimedia vertical than adults after most of the non-browsing events. We also observed that the fraction of sessions in which users aged 10 to 19 started the session with a multimedia query was 3% against 0.7% for adults. Even though these percentages are small, the differences in the proportions between age groups are still meaningful, since we are estimating these values from all type of search sessions, in which web (contrary to multimedia) search is predominant.

We observed that the likelihood of performing a multimedia search for users up to 10 years old is comparable to the likelihood found for adults. We expected a larger

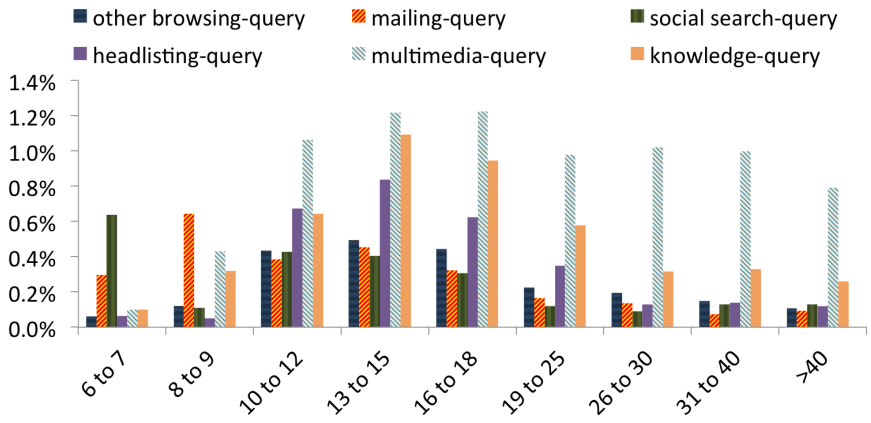


Figure 4.6: Browsing pairs likelihoods (Multimedia search).

usage of multimedia verticals for the youngest group of users given that the content available on these verticals are generally easier to parse than the websites returned from the standard web search. The low usage found for these users may also be due to the fact that users in these age ranges have been reported to have difficulties to identify the tabs and hyper-links to the non-web verticals [Druin et al., 2010].

We found that *web search* events (from the same query) had the greatest likelihood of leading to a multimedia search query. The fraction of *web query* events that led to multimedia search was 6.9% for users between 10 and 19 and 3.7% for the other age groups. On the other hand, the fraction of *web result* events (from the same query) that led to multimedia search was around 1.0% for users between 10 to 19 and 0.3% for the other groups. The higher proportion obtained for the former type of events (*web queries*) can be explained by the fact that several users access multimedia verticals by simply clicking of the tab of image or video search, procedure that automatically send the same query to the multimedia vertical. Nonetheless, as it was mentioned before, these values suggest that children under 10 years old do not use this strategy.

The proportion of web search results that led to multimedia search (with a different query) was also significantly higher on users from 10 to 19 years old. The proportion observed for these users was of 4.5% against 1.3% for the other age groups. This result is interesting because it shows that providing rich media from different genre verticals (e.g. images) can improve the web experience of children in these age ranges (e.g. through aggregated or faceted interfaces).

We observed that the fraction of each one of the browsing events that lead to multimedia search was under 1% for all the age groups, except for the *multimedia browsing* events which classify page views under similar services (e.g. *youtube.com*, *flicker.com*) as browsing activities and account for 1.5% for the young group of users. Nonetheless, we observed that young users were more likely to perform multimedia

search after carrying out browsing activities than adults, as it was the case with the search events. Figure 4.6 depicts the likelihoods of submitting a query to a multimedia search service after a page view on each one of the browsing types. These results indicate that young users employed more often multimedia search than adults. We consider that improving the accessibility of video and image verticals to young users can highly improve their search experience given that standard web search is one of the main triggers of this type of search. Aggregated search interfaces seem to be an ideal solution to provide results for these users since it eliminates the burden associated to finding the links to the verticals or to locating the vertical results.

All the statistical tests for the proportions reported in Figure 4.5 and 4.6 were proven statistical significant with p-values $< 1E-5$. The following patterns in Figure 4.6 were not found statistical significant: *multimedia-search query* for the pair 6-7/8-9 (p-value 0.601); *headlisting-search query* for the pair 6-7/8-9 (p-value 0.548) and *mailing-search query* for the pair 10-12/13-15 (p-value 0.319).

4.6 Search trigger classification

In the previous section we explored the likelihood of a user to carry out web and multimedia search given that the previous event was a browsing or a search event. However, it is not clear that the browsing or search event triggers the new search. Consider the following two examples to illustrate the cases when a browsing event does and does not trigger a search event: *(i)* a user starts the session by going to a social networking website. The user checks the inbox for new messages. Right after the user decides to perform a web search to start collecting information for one of their university duties. In this example, the browsing activity (*social browsing*) does not trigger the search event, and the search respond to a change of mood of the user. Now consider the following example: *(ii)* a 9 years old user starts the session by checking the news. The user browses to an article in the science and environment section that discusses new methods to halt rabies. The user does not what rabies is, thus the user decides to copy and paste the word into the search engine box. After understanding that rabies is an illness, the user comes back to the article. In the article the user discovers that rabies is a major issue in certain jungles in South America. The user is intrigued by this place and decides to switch to the search engine once again. This time the user types the query: *animals in south america*. In this example, the two queries submitted are triggered by the information need generated during the browsing activity. In the first case, the query is a highly frequent keyword in the text (i.e. *rabies*). In the second case the query is an information request related to the content of the webpage and it is not explicitly found as a keyword in the page.

We attempt to quantify the proportion of browsing activity that triggers search by classifying the pairs *event* $\rightarrow$ *(search) web query* into the following categories:

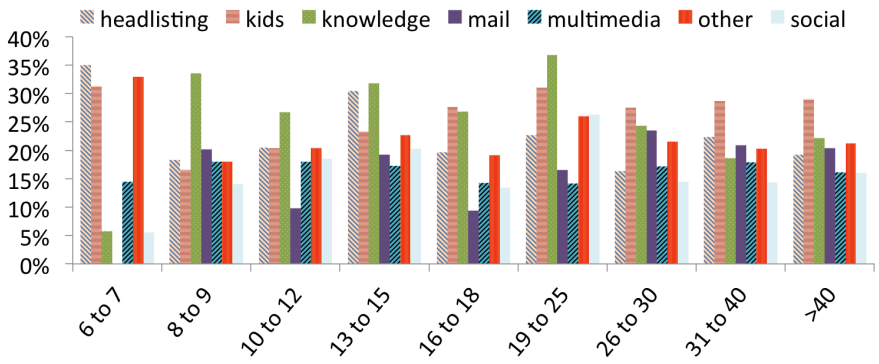


Figure 4.7: Proportion of trigger *1* for the browsing event pairs.

1. The query is explicitly mentioned in the browsing page.

2. The query is not mentioned in the browsing page but it is found in the domain of the browsing page.

3. The query is not found in either the browsing page or its domain.

The first category captures the cases in which the query submitted after the browsing event is explicitly mentioned on the webpage, and it has a high frequency. In the previous example, the submission of the query *rabies* falls under this category since the query is stated in the article several times. The second category captures the cases in which the query submitted target information that is related to the content of the browsing page explored but it not explicitly mentioned on the webpage, which was the case in the previous example for the query *animals in south america*. The third trigger occurs when the user wants to visit a well known resource but the user does not recall the exact url (e.g. *facebook*, *google*).

These categories were proposed by Cheng et al. [2010] after manually inspecting a sample of 200 sessions containing distinct browsing-to-search patterns. A set of technologies were built to detect each case automatically. For the first category, the query keywords are matched with the browsing website and a threshold is set to classify only high frequent keywords. For the second category an analogous procedure is carried out in the domain of the Website.

Figure 4.7 and Figure 4.8 depict the proportion of browsing event pairs triggered by the first category (*keyword*) for the set of non-navigational and navigational queries, respectively. The results reported were obtained by normalizing over the set of event pairs of each browsing type. The proportions were estimated on the set of browsing events that we were able to classify automatically. On overall for each one of the age groups we classified around 40% of the event pairs.

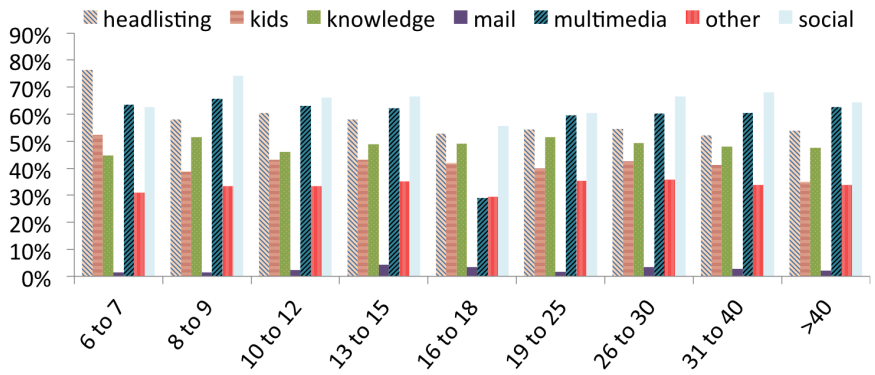


Figure 4.8: Proportion of triggers *1* and *2* for the browsing event pairs.

Particularly, we found a high proportion of the *keyword trigger category* for the knowledge browsing event throughout all the age ranges, except for users aged 6 to 7 years old. The proportion for users between 8 to 19 years old was around 0.32, and for older users was 0.23 approximately. The trend observed is consistent with the findings described in the previous section. Recall that the probability of a *knowledge* page view leading to search activity was higher for young users. This finding indicates that a high proportion of these pairs indeed correspond to search triggers and not necessarily to a change of mood or a change of topical interest. However, this trend was not observed for the youngest group. For these users the proportion of search query events triggered by keywords mentioned in the knowledge page was only 5%. This result suggest that these users do not take the same effort shown by older users to carry out follow-up searches related to the content they are browsing.

Interestingly, in Figure 4.7 is also observed that the proportion of search triggered by *knowledge search* events decays for users over 25 years old. This may indicate that these users do not need to carry out follow up searches as often as young users given their greater language and topical capabilities. As it was shown in Chapter 3 in sections 3.5.1 and 3.5.5, adult users have significantly larger query term vocabulary and greater diversity in the topics searched. This result may also be due to the greater search expertise of adults and due to the more developed criteria to self regulate their searches and identify when the information collected so far is sufficient or not.

In respect to the *mail* and *multimedia* browsing activities, we did not observe a clear differences between the age groups. For all the age ranges, the proportion of *keyword* search category was around 15% and 18% respectively.

As it was expected, the social browsing event triggered a higher proportion of search activity for the case of teenagers (13 to 19 years old), and the proportion decayed for children up to 12 years old and users over 19 years old. *Headlisting* also triggered a significant proportion of search activity for all age ranges, particularly

for teenagers. This result was expected given that users often carry out follow up searches to find information about items they are planning to purchase (e.g. review, specs pages) or they have interest in.

In terms of statistical significance, all the proportions and pairs in Figure 4.7 and 4.8 were statistical significant (once again with very small p-values ($< 1E-5$), except for the following pairs for Figure 4.7: multimedia for 8-9/10-12 (p-value 0.362), 10-12/13-15 (p-value 0.201) and 8-9/13-15 (p-value 0.045).

For Figure 4.8: knowledge for 13-15/16-18 (p-value 0.062); social for 13-15/16-18 (p-value 0.304); multimedia for 10-12/13-15 (p-value 0.421); kids for 10-12/13-15 (p-value 0.415), 13-15/16-18 (p-value 0.203) and 10-12/16-18 (p-value 0.354).

4.7 Conclusions and future work

4.7.1 Findings summary

In regard to *RQ-3.1*, a larger proportion of browsing activity (in respect to search activity) was observed for all the age groups, however we found that young users had a greater proportion of browsing events. 85%, 80% and 81% of the toolbar session entries were browsing activities for children, teenager and adults respectively. We believe this result is due to the topic bias towards gaming and social services of children and teenagers, services that are easily accessible without a search engine. This result also indicates that current search engines need to address the search difficulties that children and teenagers face in the Internet to attract more this segment of users.

For *RQ-3.2* we found that most of the users engaged search from the beginning of the session, although this behavior was more pronounced in the youngest group of users. This was also the case for teenagers in the case of multimedia search. We found that knowledge intense websites (e.g Wikipedia) had the highest likelihood of triggering search for all the age ranges among all the browsing activities. Interestingly, teenagers were found to be twice as likely to submit queries on multimedia services than other age groups. Web search was the most likely activity to trigger multimedia search. For the browsing activities knowledge and headlisting were the two type of pageviews with the largest likelihood of triggering multimedia search.

4.7.2 Recommendation for the development of IR technology for children

Aggregated interfaces are a promising alternative to present results from multimedia verticals related to the web results. We believe that this type of interfaces can greatly benefit the search experience of children and teenagers. Recall that we observed a large likelihood of query submission to these services after carrying out

browsing activities. Tools to find complementary material for educational content (e.g. Wikipedia, Simple Wikipedia) can also support the information collected by young users. These tools may involve finding related audio-visual material and tools for language simplification.

From our findings, methods to improve the access to knowledge search services (with content readable by children) and multimedia content are urgently required by the youngest group of users.

Chapter 5

Query Recommendation for Young Users

This chapter is based on Duarte Torres et al. [2012, 2014b].

5.1 Introduction

In this chapter we explore query recommendation methods to help young users construct queries with query senses targeting content suitable for them. Query recommendation functionality is commonly triggered by state-of-the art search engines when a query is submitted. It consists of providing related queries to help users refine their search and improve their chances to find what they are looking for.

The queries recommended can be composed of different words or can expand the original query by means of keywords. The former type of query recommendations are commonly elicited from search logs [Baeza-Yates and Tiberi, 2007; Wang and Davison, 2008]. The latter approach is also known in the literature as *query suggestion* and it is suited to help users focus the search and alleviate the problem of finding the right keywords for the query, which is particularly challenging for children [Druin et al., 2009].

In this work we focus on providing query recommendation by suggesting keywords or phrases that can be added to the user's query. We opted for this option because it is a natural way to help young users to expand their queries to focus the search. As we observed in Chapter 3, young users' queries are on average shorter and they have a significantly smaller vocabulary in respect to adult users. We believe that a careful selection of this type of query recommendation can provide a high impact in the search experience of young users. The methods proposed in this chapter also intend to mitigate the problem of finding irrelevant and unsuitable information, since our suggestions boost search aspects that are related to young users search intents.

Extensive research has been carried out on general-purpose query recommendation. Our work deviates from previous studies in that *(1)* the suggestions are aimed at children's and teenagers' search intents and *(2)* the suggestions are constructed in the absence of query logs.

We explore the construction of queries based on tags (from social media) that are associated to the query web results and to previously known web resources for young users. Tags from the bookmarking system *Delicious* are employed. We consider tags a reasonable source of terms for query expansion given the high overlap of tags and query terms found in large query logs [Yanbe et al., 2007]. Tags are a valuable resource in the IR domain for children and teenagers since we can exploit the collaborative information provided by users sharing web resources for this segment of users.

A novel method is proposed to boost tags that are frequently used to describe resources for young users through a biased random walk based on information metrics. The assumption of our method is that tags frequently associated to urls focused on young users topics are better candidates to construct query suggestions for this public. For instance consider the query *cars*. According to Google's query suggestions, common aspects of this query are *car rentals, cars for sale, used cars, new cars, disney cars and car pictures*. On the other hand, aspects oriented to satisfy young users information needs should rather include *car movies, car games, car toys, car coloring pages, car pictures, car crafts*. Our system ranks higher the latter tags, providing more focused suggestions on content for young users (children in this case). This method is compared against state of the art query suggestions and against other biased random walks to prove its effectiveness. This can be stated as a research question in the following way:

R.Q-5.1: To what extent does a random walk, biased by using information gain metrics, improve the effectiveness of the query recommendations for young users over traditional biased and unbiased random walks?

In the second part of the chapter is explored the usage of a learning to rank approach to improve the ranking of the query suggestions. We employ the random walk score along language models and topic features based on the category structure of the Dmoz *kids and teens* section, in this way we emphasize further the suitability of the suggestions for these users and the system is informed about the relevant topics associated to the query. We explore the benefits and limitation on this approach:

R.Q-5.2: Can we improve the quality of the ranking of query recommendations by combining the random walk score with features based on language models and topical knowledge?

The quality of the results is evaluated using a large anonymized log sample of queries and query reformulations from users aged 8 to 18 years old extracted from

the Yahoo! Search engine. Additionally, we also employ a selected set of query reformulations used to retrieve content aimed at children from the AOL query log. The motivation of using the two test sets is to explore at what extend is possible to use queries targeting content for young users, but not necessarily submitted by them, for the evaluation of query recommendation methods in a large scale. This is a question worth to be explored given the difficulty to access a large-scale set of queries submitted by very young users:

R.Q-5.3: Can we substitute successful query reformulations submitted by young users by using query reformulations targeting content for this user segment in the evaluation of query recommendation?

The chapter is organized in the following manner: Section 5.2 discusses the relevant related work to this chapter. Section 5.3 describes our random walk method and the query representation employed. Section 5.4 describes in detailed how three well established biased random walk can be mapped to our problem settings. Section 5.5 describes the data acquisition process. Section 5.6 presents the results obtained by our random walk method by using the two test sets (Yahoo! Search and AOL logs). We contrast our results against the selected biased random walks and state-of-the art search engine query suggestions. Section 5.7 presents the features utilized to improve further the ranking of results and the experimental results. In the last section is summarized the findings in respect to the research questions and directions for future work are discussed.

5.2 Related work

This work is related to the areas of query suggestion, query expansion, tag ranking and biased random walks.

5.2.1 Query Recommendation

Extensive research has been carried out on query recommendation based on click-through data from query logs [Baeza-Yates and Tiberi, 2007; Gao et al., 2010]. In these methods the association between query and documents in the search graph is mined to infer related queries. More recently, random walk frameworks have been proposed to rank documents and queries using hitting time [Mei et al., 2008] and based on the query-document frequency in the graph [Boldi et al., 2009; Craswell and Szummer, 2007].

The method we propose utilizes the framework proposed by Craswell and Szummer [2007]. However, our work deviates from theirs in the definition of the transition

probabilities and the normalization of these probabilities. Our motivation is to bias the walk towards suggestions more appropriate for a specific niche of users (i.e. children) which is not addressed by their work.

Recently, two random walk frameworks have been proposed to leverage query log and social media annotation within the same graph. The first exploits the latent topic space of the graph [Bing et al., 2011] and the second utilizes the graph hitting time [Mei et al., 2008]. Their focus is to refine queries by exploiting the tag vocabulary of social media and to provide exploratory and search query suggestions within the same framework. Our work also exploits the annotations of social media for the generation of query suggestions. However, our work differs in that the query suggestions are generated in the absence of query logs, that is solely based on social media. Their work also does not address the generation of suggestions for a specific group of users, which we address by introducing a novel bias into the random walk. Finally, our work differs in that we integrate the scores of the random walk in a learning to rank approach to emphasize further the suitability of the suggestions for children.

5.2.2 IR for Children

Gyllstrom and Moens [2010] presented a variation of Page Rank to rank web pages that are more suitable for children. They utilized label propagation to score documents. Our work deviates in the characteristics of the graph utilized (we employ social media while they employ web documents) and in that we boost query suggestions associated with content for children through information metrics between a foreground model of tags used to describe content for children and a background model. Moreover we improve over the results of our random walk using a learning to rank framework by introducing features that are not trivial to add in a graph based model. Gyllstrom and Moens [2011] presented a method to detect queries that represent controversial topics and topics for children. Their method filter topics by relying on the query suggestions of a commercial search engine. In our work we provide query suggestions in the absence of search engine's query suggestion functionality. Eickhoff et al. [2010] proposed a machine learning approach to filter content unsuitable for children. Although both problems are different (they addressed a binary classification problem), some of the features we use are similar to the ones used in their research. Nonetheless, the main feature in our approach is the random walk proposed.

It is also worth to mention other efforts from the PuppyIR European project[1] in which some the content filtering, query suggestion, and topic detection mentioned has been applied and presented as showcases [Azzopardi et al., 2012a,b,c; van der Sluis et al., 2011].

[1] `http://puppyir.eu`

5.2.3 Tag Ranking

Tag ranking has recently received attention given the proliferation of social media sharing sites. Liu et al. [2009] proposed a method to estimate the relevance score of a tag to an image based on probability density estimation. The estimation is further refined using a random walk over a tag similarity graph. Several tag ranking methods have been proposed in the domain of image tagging [Feng et al., 2012; Li et al., 2012, 2008; Zhu et al., 2010; Zhuang and Hoi, 2011]. On overall, these methods exploit the similarity between tags based on the neighboring tags in the graph. Additional evidence such as visual and semantic features and are also employed to reduce noise and disambiguate the meaning of the tags.

Our work deviates in the structure of the graph and the bias introduced into the random walk. In the previous works mentioned the graph consists only of tags and in our problem it consists of tags and web resources. This graph structure is important since we exploit the characteristics of web resources aimed at young users to bias the random walk. Moreover, they do not consider the age dimension of the users.

5.2.4 Biased random walks

Biased random walks have been proposed before in different problem domains. Haveliwala [2002, 2003] proposed the Topical Page Rank. This is a variation of PageRank in which the topic of the query and the urls are taken into account. The PageRank score is refined by calculating multiple scores for each page, each score representing the importance of the page in respect to a topic. These scores are then combined according the importance of each query for the topic. The performance of this method is further explored by Kohlschütter et al. [2007]. They found that the performance of this method varies according the specificity of the topics considered, and that more specific topics tend to lead to better results. Qiu and Cho build on the Topical Page Rank and includes the user clicking story to enhance the ranking. Abou-Assaleh et al. [2007] presents a similar method in which the search is focused on specific topics. They obtained comparable results with less computational overhead.

Wu and Chellapilla [2007] proposed a biased random walk to extract link spam communities when at least one of the members of this community is known. Their method is referred as Spam Rank. The bias is introduced by employing decay probabilities, in which nodes having a greater distance from the seed set are penalized. Zhang et al. [2009] expanded Wu and Chellapilla [2007]'s work by proposing a method to automatically enlarge the seed set. Gyöngyi et al. [2004] addressed the problem of ranking pages avoiding spam. They used a seed set of trusted resources to avoid spam instead of having a seed set to detect spam communities.

Fuxman et al. [2008] presented a random walk with absorbing states for the generation of keywords in the domain of sponsored search. The random walk is defined

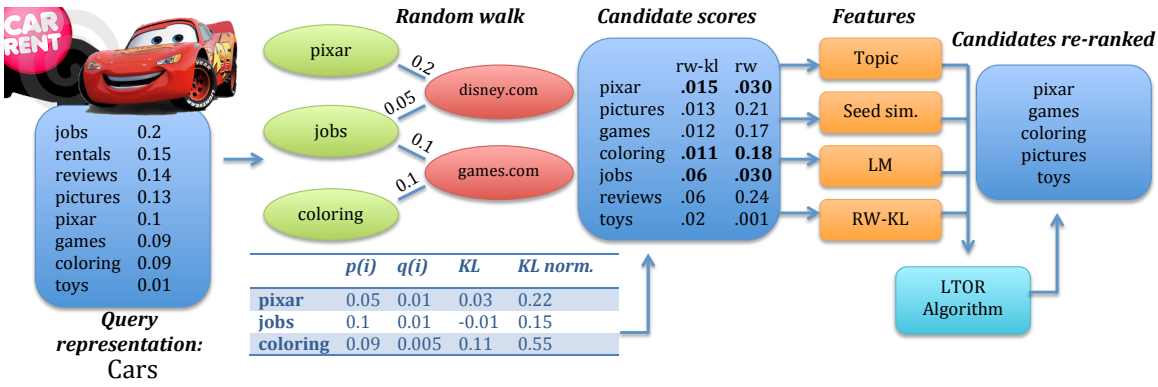


Figure 5.1: Query Suggestions Framework using the query *cars* as an example.

on a bipartite graph of queries and urls constructed from query logs. A set of seed queries or urls are set as absorbing states to bias the random walk towards these states and the relationship between queries and urls in the graph are exploited to generate keyword suggestions. Their approach relates to our problem in that we are interested in biased the random walk from a seed set of urls and tags. The generation of keywords in this domain has been explored further [Hui et al., 2013; Ravi et al., 2010].

In section 5.4 we present a more detailed description of the three methods: Topical Page Rank, Spam Rank and the generation of keywords proposed by Fuxman et al. [2008]. We describe details of how these methods can be mapped to the scenario addressed in this chapter.

5.3 Method

We envisage a search service which uses state-of-the art search engines to deliver content aimed at children [Duarte Torres, 2011]. In this system, the query submitted by the user is sent to several search engines to retrieve keywords from the snippets and titles of the web results. These keywords represent the possible topics associated to the user's query. Our task is to generate these keywords and rank them to construct query suggestions. The ranking is carried out by first generating a ranked list of candidate suggestions using the random walk. Secondly, the candidates' ranking is refined with a learning to rank approach, in which the random walk score is combined with topic features and language model features. Figure 5.1 depicts the framework employed to rank query suggestions. Note that under this scenario we do not have access to search engine query logs which are widely used for query recommendation [Boldi et al., 2009; Ma et al., 2011]. Although in our architecture it would be possible

to register log activity, we would still face the cold start problems of not having data to generate the query suggestions. Moreover given the increasing privacy concerns and the characteristics of the audience targeted by our system (i.e. children), it is desirable to avoid gathering user information. Recent search engines as *DuckDuckGo*[1] and *Yippy*[2] are gaining popularity in part for their policy of not storing user data.

5.3.1 Random Walk Towards Content for Children

Our random walk model uses a bipartite graph of web resources (i.e. *urls*) and tag nodes. Previous research on tag ranking [Liu et al., 2009] employed random walks methods for tag recommendation systems using a graph composed solely of tags. In the setting of our problem is useful to treat urls as nodes as well, since our method rely on a trusted set of web resources, which are used as seeds to bias the random walk towards more relevant tags for the targeted audience. That is, tags more frequently associated to urls that are known to be used by young users will be promoted over tags employed more frequently to described urls for a different niche of users (i.e. adults). Note that it is not straightforward to represent this information in graphs that are only composed of tag nodes, moreover in this graph representation is possible to add a measure of how reliable or trustful a seed url is.

In this work the graph was created using a set of the *Del.icio.us* bookmarks from the collection presented by Wetzker et al. [2008]. Concretely, bookmarks of urls known to be adequate for children and young audiences were extracted to create the set of urls and tags. Details about the characteristics of the dataset are provided in Section 5.5.1. Our random walk method is based on the framework proposed by Craswell and Szummer [2007]. Formally the graph is defined as:

Definition 1. (bipartite graph of urls and tags)*:* $G = \langle U, T, E \rangle$
where $E = \{(u,t)|(u,t) \in U \times T\}$*,* $U = \{u_1, u_2, .., u_n\}$ *is the set of urls described by tags* $T = \{t_1, t_2, .., t_m\}$ *and* E *is the set of edges in the graph.*

Craswell and Szummer [2007] defines the transition probabilities as:

$$p_{fw}(i|j) = \begin{cases} (1-\alpha)\frac{c(i,j)}{\sum_{k:(j,k)\in E} c(j,k)} & if\ i \neq j \\ \alpha & if\ i = j \end{cases} \tag{5.1}$$

$$p_{bw}(i|j) = \begin{cases} (1-\alpha)\frac{p_{fw}(j|i)}{\sum_{k:(i,k)\in E} p_{fw}(k|i)} & if\ i \neq j \\ \alpha & if\ i = j \end{cases} \tag{5.2}$$

For the cases of forward and backward random walk respectively. The term $c(i,j)$ represents the number of times a tag i was used to describe a web resource j and

[1] https://duckduckgo.com/privacy.htm
[2] http://search.yippy.com/privacy

the term α is the self transition probability which is used to slow the diffusion of the scores. We employed both types of weighted functions as baselines. As it was observed by Boldi et al. [2009], we found that the backward weight performs better, although only marginally for our problem, as will be shown in the next section.

We propose to bias the random walk by introducing a weight based on the point-wise Kullback-Leibler (KL) divergence metric. Intuitively, this metric promotes those tags that have a greater likelihood to appear in a collection of content for children (our foreground model) than in a corpus of content for adults (background model). This intuition is exemplified in Figure 5.1 using the query *cars*. In this example the most popular keywords associated to the query are *jobs*, *rentals* and *reviews*. Using the baseline random walk we obtain *pixar*, *jobs* and *reviews* as the top ranked results. However, when the *KL* weight is introduced, the latter two keywords are penalized further allowing keywords such as *pictures*, *games* and *coloring* to be ranked better. Equation 5.3 and 5.4 reflect the new transition functions.

$$p_{fwKL}(i|j) = p(i) \log \frac{p(j)}{g(j)} p_{fw}(i|j) \tag{5.3}$$

$$p_{bwKL}(i|j) = \left\{ \begin{array}{cc} (1-\alpha)\frac{p_{fwKL}(j|i)}{\sum_{k:(i,k)\in E} p_{fwKL}(k|i)} & if\ i \neq j \\ \alpha & if\ i = j \end{array} \right\} \tag{5.4}$$

where $p(i)$ is the probability of a tag (or url) to appear in the collection of resources for children and $g(j)$ is the probability of i to appear in the collection of resources for the general public. We normalize the point-wise Kullback-Leibler (KL) distances to lie between 0 and 1 in order to introduce them into the random walk framework. The normalization was carried out using the maximum and minimum KL point-wise distance in the collection in the following manner: $kl_n(p||q) = (kl(p||q) - \min(KL))/(\max(KL) - \min(KL))$.

We also found that using a uniform normalization for the transition of *urls* to *tags* improves the performance of the random walk. Intuitively, this occurs because the standard transitions of urls to tags tend to promote the most popular tags. However, our focus is to promote tags that are more children (or teenager) oriented, which are not necessarily the most popular for a given url. Thus, a uniform normalization emphasizes the effect of the KL weight introduced in Equation 5.3 and 5.4. Using this observation we renormalized the forward probability as follows:

$$p_{fwN}(i|j) = \left\{ \begin{array}{cc} (1-\alpha)\frac{c(i,j)}{\sum_{k:(j,k)\in E} c(j,k)} & if\ i \neq j,\ j \in T \\ (1-\alpha)\frac{p_{fw}(j|i)}{\sum_{i:(j,i)\in E} p_{fw}(j|i)} & if\ i \neq j,\ j \in U \\ \alpha & if\ i = j \end{array} \right\} \tag{5.5}$$

From Equation 5.3, we need to estimate the probabilities of the tags and urls

in the two corpora. These probabilities are estimated based on a set of Delicious bookmarks that represent the interests of the target group.

We define a bookmark as a tuple containing an url and a tag which describes the url: $b = \langle u_i, t_i \rangle$ where $u_i \in U, t_i \in T$, the set of urls and tags respectively. A collection of bookmarks is defined as a bag of N bookmarks $B = \{1, b_2, .., b_N\}$.

We employ a set of bookmarks that contains trusted urls oriented towards a specific audience: children and teenagers.

Definition 2. *The* (bag of bookmarks of trusted and oriented urls for a target audience) *is defined as:*
$B_k = \{b_1, b_2, .., b_N | url(b_i) \in U_k\}$ *where* U_k *is the set of seeds urls and* $url(b_i)$ *extracts the url from the bookmark* b_i*.*

The estimation of the transition probabilities depicted in Equation 5.3 is estimated using maximum-likelihood estimation (MLE) using B_k for the foreground model and B for the background model.

$$\begin{aligned} p(t) = \frac{cf_{B_k}(t)}{|T|}, p(u) = \frac{cf_{B_k}(u)}{|U|} \\ g(t) = \frac{cf_{B}(t)}{|T|}, g(u) = \frac{cf_{B}(u)}{|U|} \end{aligned} \tag{5.6}$$

where $|T|$ and $|U|$ are the raw size of tags and urls in the collection B_k.

5.3.2 Query Representation

The query is represented as a single node in the graph and we define a special transition probability from the query node to the tag nodes of the graph. We do not include transition probabilities from the query to url nodes because the user's query is represented as a bag of tags. The query representation is constructed from the query itself and the tags found in the titles and snippets of the top ranked web results. The query can also be seen as a document constructed with the tags found on the web results and the query. Formally we define the user's query and the tag set of a query as:

Definition 3. (User's query) *The user's query of length* l *is represented as the sequence of words* $(w_1, w_2, .., w_l that are known in the vocabulary of tags)$*.*

Definition 4. (Tag set of a query) *The tag set of a query* q *consists of the* m *tags extracted from a social bookmarking system* S*, which are associated to the top web results of query* q*:* $Q = \{t_1, t_2, .., t_m\}$*.*

Thus, the query q is the union of the the set Q and the tokens that make part of the user's query and that are known in the vocabulary of tags. This representation

is convenient because query suggestions can often be obtained directly from the keywords appearing in the snippets of the web results [Yanbe et al., 2007]. Using this query representation we define the transition probability $p(t|q)$ as:

$$\begin{aligned} p(t|q) &= \frac{p(q|t)p(t)}{p(q)} \\ p(t|q) &\propto p(t)p(q|t) \\ p(t|q) &\propto p(t)\prod_{i=1}^{|q|} p(q_i|t) \end{aligned} \tag{5.7}$$

The first term on the right hand side is the likelihood of the candidate tag t in the collection and the second term describes the likelihood of t co-occurring between the tags in the query and the collection. These probabilities are estimated using MLE in a similar fashion as in 5.6

$$p(q_i|t) = \frac{cf(q_i,t) + \mu\, p(q_i)}{|T| + \mu} \tag{5.8}$$

where $p(q_i)$ is the prior probability of q_i and μ is the Dirichtlet smoothing parameter.

5.4 Related biased random walks

In the following paragraphs we present the description of other biased random walks that can be applied to our problem. We point out the differences between these methods and our approach and we provide relevant implementation details adopted for their comparison against our method.

5.4.1 Topic-sensitive page rank

The topic-sensitive page rank proposed by Haveliwala [2002] builds on the definition of PageRank, in which the scores of the nodes are computed in the following manner:

$$\vec{Score}_{t+1} = (1-\alpha)M \times \vec{Score}_t + \alpha\vec{p} \tag{5.9}$$

where the element m_ij of matrix M is equal to $1/N_j$ (the inverse of the number of outgoing links of node j) if there is a link from node j to node i and 0 otherwise. And $\vec{p} = \left[\frac{1}{|N|}\right]_{N\times 1}$.

Thus, the prior probabilities are uniform for all the nodes and the transition probabilities are normalized uniformly by the number of outgoing edges. The bias is

introduced by using as vector $\vec{p}$ the topic vector $\vec{v}$ (i.e. $\vec{p}$= $\vec{v}$), in which each element in this new vector is defined in the following manner:

$$v_i = \begin{cases} \frac{1}{|T_j|} & if\ i \in T \\ 0 & if\ i \notin T \end{cases} \tag{5.10}$$

Haveliwala [2002] defines the query representation based on the probability of the query given the target topic:

$$p(c_j|q) = \frac{p(c_j)p(q|c_j)}{p(q)} \propto p(c_j) \prod_i p(q_i|c_j) \tag{5.11}$$

where $p(c_j)$ is the probability of topic c_j (set uniformly in [Haveliwala, 2002]) and $p(q_i|c_j)$ is estimated using MLE based on the documents of the Dmoz category c_j. The scores of all the topics are combined in the following manner:

$$s_{qn} = \sum_j p(c_j|q) score_{jn} \tag{5.12}$$

where $score_{qn}$ is the score of node n in the graph (e.g. a url in PageRank) for the query q.

The first key difference of the topic-sensitive PageRank method with the method proposed in our work is that all the nodes are considered of the same class, while we propose a graph that distinguishes between tag and url nodes. Also note that the main bias introduced by this method arises in the non-uniform probabilities of vector $\vec{v}$, which makes uses of the number of outgoing links belonging to the target topic. This bias is implicit in our method by the construction of the graph in which only urls and tags from trusted resources for children are utilized. We implement the topic-sensitive PageRank by considering that we are only targeting topics for children (or teenagers). Thus we do not employ a vector of scores of different topics, instead we estimate the scores for the children set of urls found in Dmoz as we do for our method. We also utilize the query definition proposed in Equation 5.8 to make the results comparable to our method.

5.4.2 Seed based random walk

Fuxman et al. [2008] proposed a random walk using a bipartite graph of urls and queries (tags in our problem scenario) from query logs. Their algorithm assumes as input the graph, a set of concepts and a set of seeds of urls or tags in which a seed is mapped to a specific class. Their task is to recommend urls or tags related to the seed set representing the concepts. We map their model to our settings by employing only one class (*i.e.* content for children/teenagers) and by employing the bag of tags

that represent the user's query as the set of seeds. The task in our problem is to find other tags related to the seed tags. The transition probabilities of this random walk are defined in the following fashion [Fuxman et al., 2008]:

$$p(l_t = c) = (1 - \alpha) \sum_{u:(t,u)\in E} w_{tu} p(l_u = c) \tag{5.13}$$

where the weight values $w_{tu} = \frac{c(t,u)}{\sum_{k:(t,k)\in E} c(t,k)}$ are the transition probabilities from node t to node u. The value $p(l_t = c)$ represents the probability of a tag or url of being absorbed by the nodes in the set seed during the random walk. In practice these values are based on the accumulated random walk score. The transaction probability from node u to node t is defined analogously.

Note that these transition probabilities are equivalent to the transition probabilities defined in Equation 5.1. The main difference between the method proposed by Fuxman et al. [2008] and our method is the way the bias is introduced to the random walk. Fuxman et al. [2008] introduces the bias by establishing all the nodes from the seed set (either urls or tags) as absorbing states. This is accomplished by setting $p(l_t) = 1$ if the node t belongs to the seed set and $p(l_t) = 0$ otherwise. Note that this implies that the values of the nodes belonging to the seed set are not updated with the random walk, since they are set to 1 in the initialization step. Another difference is the special normalization scheme adopted for the random walk (Equation 5.5) and the use of a special query node to represent the query tags. Additionally, a threshold is employed by Fuxman et al. [2008] to set as null absorbing states (nodes with $p(l_t) = 0$) those nodes in which the accumulated probability fall under the threshold. This threshold is also used for efficiency purposes in the implementation.

5.4.3 Spam detection random walk

Wu and Chellapilla [2007] employed a biased random walk to extract spam communities. The input of the algorithm is a graph (all the nodes are of the same type) and a seed set of nodes, which are used to bias the random walk. We map this method to our problem scenario by employing the query bag of tags as the seed set. In this case the tags related to the seed tags are seen as the spam community to be detected.

The node probabilities are updated using the following expression:

$$p(i)_{t+1} = \frac{1}{2} \left[I + AD \right] p_t(i) \tag{5.14}$$

where I is the identity matrix, A is the adjacency matrix of the graph and D is the diagonal matrix in which the elements $d_{ii} = \frac{1}{d(i)}$ where $d(i)$ is the degree of node i. The bias is introduced in the initialization step of the random walk, by setting the

node probabilities in the following fashion:

$$p(i)_t = \begin{cases} 1/|S| & if\ i \in S \\ 0 & otherwise \end{cases} \tag{5.15}$$

where S is set of seed nodes and $|S|$ is the size of the set.

Wu and Chellapilla [2007] added four additional constraints to the random walk in the implementation: *(i)* The probability scores are truncated to zero if they fall under a specified threshold or if they fall in the bottom of the k-percentile probability distribution. We adopted the former in our implementation. Note that an analogous parameter is utilized in Fuxman et al. [2008] for optimization purposes; *(ii)* the probabilities are renormalized to sum to one after each random walk iteration. This process is carried out if there are leaf nodes with no children nodes. However, this situation does not occur given the construction of our bipartite graph; *(iii)* a list of trusted (*white*) domains are considered to prevent the random walk from following links to these set of domains. This restriction is reasonable for spam detection since trusted domains are unlikely to link to spam. However, we believe this restriction is specific for spam detection and we did not consider it in our problem scenario; and *(iv)* the node probabilities are biased by penalizing those nodes that have a greater distance in respect to the seed nodes. This is carried out by weighting the node probability $p(i)_t = p(i)_t \cdot 2^{\delta}(i)$, where $\delta(i)$ is the shortest path distance of node i to the nodes in the seed set. This restriction was considered.

This method differs from our method in that it does not distinguish the types of nodes (as it was the case with the topic-sensitive PageRank) and the transition probabilities are defined based on the number of nodes in the seed set, thus the information about the relationship of tags and urls is not captured. As it was the case with the previous methods, the information of suitability for children represented through the point-wise Kullback Leibler divergence metric is not captured.

5.5 Data set extraction

5.5.1 Training Data

As training data we created a set of *Del.icio.us* bookmarks from the collection created by Wetzker et al. [2008]. To the best of our knowledge this is the largest collection of social tagged data available for research. The collection contains 132 million bookmarks and 420 million tag assignments and was retrieved between December 2007 and April 2008. The set was created by extracting the bookmarks of the urls listed in the *Kids and Teens* section of *Dmoz*. These urls link to "web sites that have been

selected for age-appropriate content by a team of volunteer editors"[1]. These resources have also been used in other information retrieval problems for young users [Eickhoff et al., 2010; Gyllstrom and Moens, 2010] with positive results.

The data cleaning process has a particular importance in our problem since we require well-formed and meaningful tags to construct query suggestions. We observed that tags are noisy and their usage varies greatly among users. We estimated that 9% of the tag volume were not useful as candidates for query expansion, either because the tags were not descriptive or because they referred to web addresses (e.g. *to-see, www.sfgate.com*). We also found a high percentage of ill-formed descriptive tags (e.g. *artist (music)*) and a large percentage of multi-worded tags: both with and without token separators (e.g. *new-york, avrillavigne*). Traditionally these problems are addressed relying on the redundancy of the data. However, in our problem other strategies are required given that the volume of information aimed at children (and in general to a niche of users) is significantly smaller than the volume of data aimed at the average user.

The data cleaning process was carried out in two steps: *tag normalization* and *tag filtering*. For the normalization we first follow a rule-based approach to generate a homogeneous representation of the multi-worded tags. Token separators such as "_", ".", " " were normalized to the character "_". For example this procedure maps the tags *star.wars* and *star-stars* to *star_wars* To normalize multi-worded tags (e.g. *avrillavigne*) we define a relation R between the set of tags without token separation T and the set of known multi-worded tags in the collection MT, in which each tag of the form $x_1x_2..x_n$ is associated to one or more of their split forms (if any) $x_1\_x_2\_..\_x_n$. The relation is defined as

$$R = \{(a, B)|a \in T, B \subseteq T_h, b \in T_h \wedge a = b_{untokenized}\} \tag{5.16}$$

where $b_{untokenized}$ is equal to the tag b without any token separation. This relation gives us a set of candidate split forms for a target tag. However, it is still necessary to decide when the tag has to be split since the split form of a tag is not always the correct form (it may be due to misuse language). Three features were employed to decide when tags of the form x_1x_2 should be split into $x_1\_x_2$: *(1)* normalized point-wise mutual information (*nPMI*); *(2)* the ratio between the frequency of the tag in the form $x_1\_x_2$ and the form x_1x_2 ($\frac{f_{x_1\_x_2}}{f_{x_1x_2}}$); *(3)* and the frequency of the tag in the form $x_1\_x_2$.

PMI is commonly used in NLP for mining collocations. A drawback of *PMI* is its sensitivity to the frequency of the terms involved in the calculation [Van de Cruys, 2011]. Higher *PMI* values indicate a higher association between the terms, nonetheless high values can also hold even if the two terms rarely occur in the collection.

[1] http://www.dmoz.com

For this reason we employed *nPMI* which is less sensitive to the sparsity of the data. Equation 5.17 shows how the *nPMI* is calculated for two terms. To calculate the *nPMI* for n terms we employed the total correlation information metric, which is one of the possible generalization of *PMI*. This metric measures the amount of information that is shared among a set a random variables [Van de Cruys, 2011]. Equation 5.17 and 5.18 shows its definition.

$$\begin{aligned} pmi(x_1, x_2) &= \log\left[\frac{p(x_1, x_2)}{p(x_1)p(x_2)}\right], \\ npmi(x_1, x_2) &= \frac{pmi(x_1, x_2)}{-\log\left[\max(p(x_1), p(x_2))\right]} \end{aligned} \tag{5.17}$$

$$\begin{aligned} pmi(x_1, x_2, .., x_n) &= \log\left[\frac{p(x_1, x_2, .., x_n)}{\prod_{i=1}^{n} p(x_i)}\right], \\ npmi(x_1, x_2, .., x_n) &= \frac{pmi(x_1, x_2, .., x_n)}{-\log\left[\max(p(x_1), p(x_2), .., p(x_n))\right]} \end{aligned} \tag{5.18}$$

The ratio $\frac{f_{x_1\text{-}x_2}}{f_{x_1x_2}}$ was employed as a feature to decide when to split hyphenized tokens, since in some cases the terms within a split tag have a high level of association but the correct form is as a single token (e.g. hummingbird). The threshold values for the three features were set experimentally to maximize precision on a sample of 2,000 tags without being split and their manually annotated correct split form. Concretely, by setting the parameter *(i)npmi* to 0.4, *(2)* the ratio to 0.015 and the *(3)* frequency threshold to 3 we were able to obtain a maximum precision of 87% on the sample extracted.

We also filter out tags satisfying any of the following conditions: *(i)* is in the dictionary of tags that refer to adult or explicit content; *(ii)* is used for personal administrative purposes (i.e. *to-do, to-see*); *(iii)* contains non-alphanumerical characters; *(iv)* is a url or points to a web service or *(iv)* was submitted by less than 3 users in the entire collection.

5.5.2 Test Data

We employed a large anonymized sample of search logs from the Yahoo! Search engine from May 2010 to August 2010. We only used logs from registered users that provided birth date, gender and a valid zip code. We used a subset of the age segments defined in Chapter 3:

- Readers: 8-9 years old
- Older children: 10-12 years old

- Teenagers: 13-15 years old
- Adults: over 18 years old

We left out the group of users aged 6-7 and 16-18 for simplicity. For each group of users we create a set of query tuples containing the query submitted by the user and a query reformulation which occurred within the same search session. In this work a search session is defined as a sequence of events $S = \langle e_1, e_2, .., e_n \rangle$ ordered in chronological order such that $timestamp(e_{i+1}) - timestamp(e_i) \leq 30$ minutes for every i. Each event can be either a query submission (e_i^q) or a click on a url (e_i^c). The time window length of 30 minutes is widely used in the literature [Huang and Efthimiadis, 2009; Jensen et al., 2006; Wang and Davison, 2008] and it has also been shown to be appropriate for search sessions of young users [Bilal, 2002; Druin et al., 2009].

A query $e_{i+1}^{q'}$ is a query reformulation of e_i^q if *(i)* the former is a prefix (e.g. *brit, britney spears*) or a suffix of the latter (e.g. *wars cheat codes, lego star wars cheat codes*), or the latter contains all the words of the former plus another word, independently of the order in which the words appears (e.g. *york giants, super bowl york giants xxv*) and *(ii)* there are no query events between them, although there can be arbitrary number of click events between the two queries (i.e. $S = \left\langle e_{i+1}^q, e_{i+2}^c, e_{i+3}^c, e_{i+4}^{q'} \right\rangle$ is allowed).

Using this procedure we obtained in the order of thousands of query tuples submitted by users aged 8 to 12 years old and hundreds of thousands for users over 12 years old.

We follow an analogous procedure to extract query tuples from the AOL search logs. Nonetheless since no user age information is provided in these logs we employed the method described in Chapter 2 to extract search sessions with clicks landing on trusted content for children. These queries are identified by matching the urls clicked in the log with the domains listed in the Dmoz *kids and teens* section. Search sessions were grouped according the target audience of the urls (i.e. *kids*, *teens* and *adults*) instead of the user age. In Dmoz, urls tagged as for *kids* represent urls appropriate for children aged 8 to 12. In this chapter, with *teenagers* we refer to content appropriate for users aged 13 to 15 years old.

From the AOL logs were extracted around 480K queries and 20K sessions. From these sessions we obtained in the order of tens of thousands of query pairs for the three age groups.

5.6 Random Walk Evaluation

Assessing the quality of query suggestions can be a hard task given that the intent of the user is rarely clear from solely the query. However, we consider that the query

query	source	query suggestion
monsters	bing	truck games, jobs, high, jam, energy
	rw+kl	inc,music, film, pixar, images
	AOL	scary
sol practice	bing	in computer, fourth grade
quizzes	rw+kl	test, learning, history, science sites
	AOL	world history, history
free jigsaw	bing	online, online for adults, play, natgeo
puzzles	rw+kl	puzzles games, play , for kids, online
	Yahoo!	play
the cake	bing	t-shirt, meme, lyrics, song
is a lie	rw+kl	portal, guide, walk-through, games
	Yahoo!	portal

Table 5.1: Query examples from the query logs.

suggestions that are frequently submitted by users of a given age range represent a good approximation of good query suggestions for this particular segment of users. A similar assumption has been adopted in previous query recommendation studies [Bing et al., 2011; Szpektor et al., 2011].

The performance of the query recommendation task was measured in terms of recall and NDCG. NDCG stands for Normalized Discounted Cumulative Gain, which is a measure of the effectiveness of IR systems in correctly ranking the results compared to an optimal ranking[1]. All the metrics are calculated based on the set of query tuples extracted and described in Section 5.5.2. Concretely, we have two datasets for testing: query tuples extracted from the AOL search logs, and query tuples extracted from the Yahoo! Search logs.

To calculate the performance scores we define the set of query pairs from the gold standard as $G = \{\langle q, q' \rangle\}$ where q' is a query reformulation of q. And the set $S_n = \{\langle q, q', r \rangle\}$ where q' is a query suggestion of q and r is the ranked position of q'. For instance, recall is calculated as $r = \frac{|S_n \cap G|}{|G|}$. The intersection between the set of query suggestions and query reformulations was performed using exact matching.

Table 5.1 presents query examples of the query suggestions provided by our method, the Bing search engine, and the gold standard for the queries extracted from the AOL and Yahoo! logs.

We compare the performance of the two variations of our method (Equation 5.3 and 5.4) for the forward and backward propagation schemes respectively) against the random walk baseline (Equation 5.2) established by Craswell and Szummer [2007] framework. Additionally, we compare the results obtained by our method against the Bing query suggestions. The evaluation was carried out using two test sets consisting

[1] `http://en.wikipedia.org/wiki/Discounted_cumulative_gain`

of query pairs extracted from the AOL logs and from the Yahoo! Search logs.

In a later stage of this work, we also compared the performance of our method with the three biased random walks described in Section 5.4 using the AOL search logs. This comparison was not carried using the Yahoo! Search logs since this set was no longer accessible to any of the authors at this stage.

For all the methods two data models were employed: *kids* and *teens*. The model *kids* utilizes the domains from the Dmoz directory labeled as suitable for children up to 12 years old. The *teens* model employs the domains labeled as appropriate for users from 13 to 15 years old. Recall that the graph is constructed based on the seed urls from Dmoz.

The graph constructed with the *kids* model contains 91.6K edges and 20K nodes (12.9K urls and 7.1K tags). The graph teens contains 1.3M edges and 258K nodes (62.7K tags and 195.4K urls). In the tables shown in this section the baseline will be referred as *rw-b* and the two variations of our method as *rw-kl-f* and *rw-kl-b* respectively.

For each test set (AOL and Yahoo! Search logs) we report two type of results: using pairs in which the reformulation land on a click and using pairs in which this click last at least 100 seconds, also referred as *long clicks* [Hassan et al., 2010b]. Long clicks have been shown to be a strong predictor of search success. It has been widely reported [Craswell et al., 2008; Hua et al., 2011] that users tend to click on top ranked results even if these results do not contain the information that users are looking for. Duarte Torres and Weber [2011] showed that this behavior is even stronger for users of young ages. The motivation behind using this restricted query set was to reduce this behavioral biased.

It is important to mention that there is a vocabulary gap between the training and test data since the datasets were extracted from different time windows. We found that the vocabulary intersection between the *children* query reformulations and the tag vocabulary was of 35% and 46% when using the *kids* and *teens* model. The intersection for *teenager* queries was slightly lower (32% and 41% respectively). Similar percentages were found in the AOL logs. These results indicate that the recall metrics obtained using these data sets are bounded to the percentages mentioned. All the results presented in the following sections were obtained by splitting the dataset in 10 folds and averaging the metric estimated (e.g. recall). The results reported were proven to be significant using the t-test with 0.01 level of confidence.

5.6.1 Experimental parameters

The parameters of our model and the biased random walks were set experimentally to maximize performance on an independent query sample extracted from the AOL log. The best settings were used to evaluate both query sets. For our method we set

the number of iterations of the random walk to *30* and we set the parameter α to *0.1*. The smoothing parameter μ of Equation 5.8 was set to *1200*.

For the topical sensitive PageRank we set α to 0.3 and we employed 20 iterations. For the method proposed by Fuxman et al. [2008] we set α to 0.1 and we employed 25 iterations. For the method presented in Wu and Chellapilla [2007] we set the number of iterations to 25. The parameters found are similar to the ones reported by Haveliwala [2002], Fuxman et al. [2008] and Wu and Chellapilla [2007]. We will refer to these methods in the AOL result tables as *topicalRank*, *seedRank* and *spamRank* respectively.

5.6.2 AOL Query Log Results

Tables 5.2 shows the recall values obtained for the query pairs extracted from the AOL query log for the case in of the reformulations that led to clicks. We found that both variations of our method outperform the baseline and the Bing query suggestions for the *children* and the *teenager* queries. However, this is not the case for the set of *adults* queries, which was expected given that our random walk method give priority to tags that are more popular for young users.

We observed that the maximum gain obtained is for the *children* queries when considering the top 10 suggestions: 8.0% in respect to the baseline and 10.0% in respect to Bing. For the teenagers queries the maximum gain is also obtained when considering the top 10 suggestions with the *teens* model: 6.2% in respect to the baseline and 8.1% in respect to Bing.

We also found that for all the results the best performing model is aligned with the queries they target. For instance, the *kids* models perform better on queries targeting content for children and similarly for the *teens* model. This result is interesting because the kids model is considerably smaller than the *teens* models and yet the recall scores are higher. This result suggests that simply adding web resources to the model may not lead to better results. Thus, the usefulness of the web resources seems to play an important role when ranking query suggestions.

From Table 5.2 we also observed that *rw-kl-b* performs consistently better than *rw-kl-f* and the performance difference in terms of recall between the two methods is up to 2.1%. The performance trends observed for the recall results were also reflected for the NDCG scores, as is shown in Table 5.3, which shows that the quality of the ranking is also improved by a reasonable margin.

It can be argued that these results are biased since the same collection of domains is utilized to extract the set of query pairs and to extract the set of bookmarks employed to construct the random walk graph (i.e. build the model). We show in the next section that all trends and results for the AOL logs also hold on the large set of query pairs extracted from Yahoo! Search.

query set	model	Bing	seedRank	TopicalRank	spamRank	rw-b	rw-k-f	rw-k-b	gain	gain-bias
					Top 5					
children	*kids*	1.5%	3.7%	2.3%	4.9%	3.3%	9.2%	9.5%	6.2%	4.6%
	teens		3.7%	2.3%	2.3%	2.4%	6.8%	7.4%	5.0%	5.1%
teenagers	*kids*	2.3%	3.7%	1.0%	2.8%	2.0%	5.2%	5.5%	3.5%	2.7%
	teens		4.3%	1.7%	3.9%	2.9%	8.0%	8.8%	5.9%	4.9%
adults	*kids*	5.1%	0.1%	0.1%	0.0%	0.3%	0.3%	0.4%	0.1%	0.4%
	teens		0.1%	0.1%	0.0%	0.4%	0.4%	0.4%	0.0%	0.4%
					Top 10					
children	*kids*	2.15	7.6%	2.7%	7.6%	4.1%	11.7%	12.1%	8.0%	4.5%
	teens		7.8%	3.5%	5.1%	4.5%	8.8%	9.7%	5.2%	4.6%
teenagers	*kids*	3.2%	6.3%	1.5%	3.9%	4.2%	6.5%	7.4%	3.2%	3.5%
	teens		7.2%	2.4%	4.2%	5.1%	10.5%	11.3%	6.2%	7.1%
adults	*kids*	5.1%	0.1%	0.1%	0.2%	0.7%	0.7%	0.7%	0.0%	0.5%
	teens		0.3%	0.2%	0.2%	0.9%	0.8%	0.8%	-0.1%	0.6%
					Top 50					
children	*kids*	2.1%	15.0%	9.6%	14.8%	13.2%	15.8%	16.6%	3.4%	1.8%
	teens		12.7%	8.2%	10.0%	6.9%	12.1%	12.4%	5.5%	2.4%
teenagers	*kids*	3.2%	8.1%	4.3%	8.1%	6.3%	9.9%	9.6%	3.3%	1.5%
	teens		9.6%	7.0%	11.3%	10.1%	12.7%	13.6%	3.5%	2.3%
adults	*kids*	5.1%	0.6%	0.5%	0.7%	1.0%	1.0%	1.0%	0.0%	0.3%
	teens		0.7%	0.8%	0.8%	1.3%	1.2%	1.3%	-0.1%	0.4%

Table 5.2: Recall comparison using the AOL log. Underlined values were not statistical significant when comparing our two random walks. The column *gain* expresses the performance difference between *rw-k-b* and *rw-b*. The column *gain-bias* refers to the difference between *rw-k-b* and *spamRank*.

query set	model	Bing	seedRank	TopicalRank	spamRank	rw-b	rw-k-f	rw-k-b	gain	gain-bias
Top 5										
children	*kids*	0.017	0.028	0.025	0.037	0.021	0.069	0.071	0.050	0.034
	teens		0.027	0.031	0.040	0.018	0.042	0.045	0.027	0.005
teenagers	*kids*	0.024	0.029	0.013	0.030	0.016	0.042	0.048	0.032	0.018
	teens		0.035	0.019	0.021	0.031	0.064	0.067	0.036	0.046
Top 10										
children	*kids*	0.026	0.048	0.032	0.053	0.032	0.082	0.086	0.054	0.033
	teens		0.049	0.039	0.051	0.021	0.053	0.055	0.034	0.004
teenagers	*kids*	0.031	0.042	0.016	0.036	0.017	0.045	0.046	0.029	0.010
	teens		0.049	0.022	0.029	0.036	0.075	0.078	0.042	0.049
Top 50										
children	*kids*	0.026	0.081	0.052	0.076	0.076	0.089	0.091	0.015	0.015
	teens		0.069	0.055	0.068	0.042	0.069	0.071	0.029	0.003
teenagers	*kids*	0.031	0.049	0.023	0.047	0.029	0.045	0.050	0.021	0.003
	teens		0.059	0.034	0.049	0.069	0.076	0.081	0.012	0.032

Table 5.3: NDCG comparison using the AOL search logs.

query set	model	Bing	seedRank	TopicalRank	spamRank	rw-b	rw-k-f	rw-k-b	gain	gain-bias
				Top 5						
children	*kids*	1.3%	3.6%	2.0%	3.3%	2.7%	8.9%	9.5%	6.7%	6.2%
	teens		3.3%	2.0%	5.3%	2.1%	6.6%	7.2%	5.1%	1.9%
teenagers	*kids*	2.3%	3.6%	1.2%	2.6%	1.5%	5.1%	5.5%	4.0%	2.9%
	teens		4.2%	1.9%	4.5%	2.4%	7.8%	8.8%	6.4%	4.3%
adults	*kids*	4.7%	0.0%	0.0%	0.0%	0.1%	0.3%	0.4%	0.3%	0.4%
	teens		0.0%	0.0%	0.0%	0.2%	0.3%	0.4%	0.3%	0.4%
				Top 10						
children	*kids*	1.8%	5.3%	2.6%	5.9%	3.6%	11.4%	12.0%	8.5%	6.1%
	teens		5.9%	2.6%	6.6%	4.2%	8.9%	9.5%	5.2%	2.9%
teenagers	*kids*	2.3%	5.6%	1.6%	3.5%	3.6%	6.4%	7.4%	3.8%	3.9%
	teens		6.5%	2.3%	5.8%	4.9%	10.4%	11.1%	6.3%	5.3%
adults	*kids*	4.7%	0.0%	0.0%	0.0%	0.2%	0.7%	0.7%	0.5%	0.7%
	teens		0.0%	0.3%	0.3%	0.5%	0.7%	0.7%	0.2%	0.4%
				Top 50						
children	*kids*	1.8%	13.2%	6.6%	10.5%	12.6%	15.6%	16.4%	3.8%	5.9%
	teens		11.8%	6.6%	13.2%	6.3%	12.1%	12.3%	6.0%	-0.9%
teenagers	*kids*	2.3%	7.8%	3.6%	7.8%	6.0%	9.8%	9.4%	3.4%	1.6%
	teens		9.0%	6.0%	12.4%	9.8%	12.5%	13.5%	3.7%	1.1%
adults	*kids*	4.7%	0.3%	0.0%	0.3%	0.6%	0.9%	1.0%	0.4%	0.7%
	teens		0.7%	0.7%	1.0%	0.7%	0.9%	1.3%	0.5%	0.3%

Table 5.4: Recall comparison using the AOL search logs (long clicks).

query set	model	Bing	seedRank	TopicalRank	spamRank	rw-b	rw-k-f	rw-k-b	gain	gain-bias
Top 5										
children	*kids*	0.015	0.023	0.030	0.039	0.020	0.068	0.070	0.050	0.031
	teens		0.018	0.033	0.041	0.016	0.042	0.045	0.029	0.004
teenagers	*kids*	0.017	0.028	0.013	0.024	0.016	0.042	0.045	0.029	0.021
	teens		0.032	0.017	0.031	0.029	0.063	0.067	0.038	0.036
Top 10										
children	*kids*	0.026	0.033	0.039	0.049	0.028	0.082	0.085	0.057	0.036
	teens		0.037	0.042	0.046	0.018	0.052	0.054	0.036	0.008
teenagers	*kids*	0.024	0.037	0.015	0.028	0.016	0.042	0.047	0.031	0.019
	teens		0.041	0.019	0.038	0.033	0.073	0.076	0.043	0.038
Top 50										
children	*kids*	0.026	0.059	0.047	0.061	0.076	0.088	0.090	0.015	0.029
	teens		0.055	0.050	0.064	0.040	0.068	0.071	0.031	0.007
teenagers	*kids*	0.024	0.043	0.019	0.039	0.027	0.044	0.050	0.022	0.011
	teens		0.049	0.028	0.052	0.064	0.076	0.080	0.015	0.028

Table 5.5: NDCG comparison using the AOL search logs (long clicks).

We verified that the results reported in Table 5.2 and 5.3 were statistical significant by applying a paired t-test at the 0.01 level of confidence between the mean differences reported for all the possible pair of methods considered. We found that the differences reported between all the methods (e.g. *Bing* vs. *rw-b*, *Bing* vs. *rw-kl-f*, *rw-b* vs. *rw-kl-f*) were statistical significant for all the results reported. However, the exception was for the set of adult queries for which the difference between the methods *rw-kl-f* and *rw-kl-b* were not statistical significant. Values that were not proven statistical significant (between our two random walk variations) are underlined in the result tables.

Tables 5.4 and 5.5 show the results obtained for the set of AOL query reformulations in which the reformulation leads to a *long* click. All the methods obtained lower performance values. For instance for the *children* queries Bing obtains a recall of 1.8% when considering the top 10 suggestions in contrast with the 2.1% obtained when using the first set of query pairs. Similarly the NDCG score obtained by Bing for the teenager set of queries (at top 10 suggestions) is of 0.031 in the first set and 0.024 in the second. These results suggest that the problem of predicting query reformulation is harder when we are targeting reformulations that lead to long clicks. It is important to mention that even though we observed lower performance values, the ratio between our method and the two baselines were larger. For instance, when considering the top 10 suggestions the performance gain of rw-k-b in respect to *rw-b* was 8.5% and 10.2% in respect to Bing. Using the first set of query reformulations the performance gains were of 8.1% and 10.0% respectively. These values indicate that our method performs better to predict query suggestions that lead to long clicks which is convenient since these query suggestions have a higher likelihood of leading to relevant information.

5.6.3 Biased random walks results

In respect to the three biased random walks we observed that the *TopicalRank* has the lowest performance. We observed a performance loss of 1% to 3% in terms of recall in respect to *rw-b* (our first baseline). On the other hand we observed that the *seedRank* and *spamRank* perform similarly although the latter outperforms the former when considering the top 5 ranked query suggestions. These two methods outperform *rw-b* varying from 1% to 3% (in terms of recall) depending of the query set and the model employed. Nonetheless our method still performs better than all the biased random walks by a significant margin. In particular we observed that our method outperforms by a larger margin the biased random walks when considering the top 5 ranked results. For instance the recall gain with respect to *spamRank* for the *kids* model and children query set was of 4.6% and 4.4% at top 5 and top 10 respectively. The precision gain in respect to *rw-b* was of 6.2% and 8.0% respectively.

From the results obtained for the query suggestions landing on long clicks reported in Tables 5.4 and 5.5 we observed, on overall, lower performance for the three biased random walks considered, as it was the case for the Bing query suggestions and the *rw-b* method. This behavior was particularly noticeable when considering recall and NDCG at top 10. For instance for the children query set and for the *kids* model, the recall for *seedRank*, *TopicalRank* and *spamRank* was 3.7%, 2.3% and 4.9% in the query set with all the clicks while in the query set with long clicks was 3.6%, 2.0% and 3.3% respectively. At top 10, the performance varied from 7.6%, 2.7% and 7.6% to 5.3%, 2.6% and 6.6% for the three biased random walks respectively.

We found that the performance gain of our method in respect to the two best performing biased random walks for children queries tend to be higher when the evaluation is carried out on the query set with reformulations landing on *long* clicks. On the other hand we observed slightly lower performance gains for the teenager query set. For instance when using the best performing models (e.g. the model *kids* for the children query set) the recall gain in respect to *spamRank* for the children query set was 6.2% at top 5 and 6.1% at top 10, while the recall gain observed for the query reformulations landing on clicks (not only long clicks) was of 4.6% and 4.5% a top 5 and top 10 respectively. For the case of the teenager query set, the gain varied from 4.3% to 4.9% for the query set on clicks and *long* clicks respectively when considering results at top 5, and 4.3% to 4.9% when considering results at top 10.

5.6.4 Yahoo! Search Engine Logs

Recall that the previous data set represent queries aimed at retrieving content for children and this dataset represent queries submitted by users for which their age can be estimated using the user profiles. Thus the results reported from this data provide a clearer picture of the performance of the methods for queries submitted by young users. Tables 5.6 and 5.7 report the recall and NDCG scores obtained by all the methods. As it was the case with the AOL query log, we observed that our random walk method outperforms the baseline and the query suggestions from Bing. Similarly *rw-kl-b* consistently performs better than *rw-kl-f*. The performance gain obtained by our method against the baseline and Bing was on the same order (around 6.1% with respect to the baseline and 9.2% with respect to Bing).

Interestingly, we found that the biggest performance gain is obtained for the youngest group of users and the gain decreases for older users. Another important difference observed with respect to the experiments on the AOL data is that the performance gain is higher when considering the top 5 suggestions and not the top 10. This is a desirable result given the importance of ranking suggestions at the top positions in the case of child users [Duarte Torres and Weber, 2011]. This result is also reflected by the low recall Bing has for the youngest group of users, particularly at top 5 (3.6% for users aged 8-9 vs 6.2% for adults). However, at top 10 the recall

query set	model	Bing	rw-b	rw-k-f	rw-k-b	gain
			Top 5			
8-9	*kids*	3.6%	5.3%	11.1%	11.9%	6.6%
	teens		4.2%	8.0%	8.1%	3.9%
10-12	*kids*	3.7%	4.4%	8.3%	10.1%	5.7%
	teens		4.3%	5.4%	5.7%	1.4%
13-15	*kids*	4.1%	1.0%	3.0%	4.3%	3.3%
	teens		5.0%	8.6%	9.4%	4.4%
adults	*kids*	6.2%	0.3%	1.1%	2.1%	1.8%
	teens		0.4%	5.0%	5.8%	5.4%
			Top 10			
8-9	*kids*	7.6%	9.2%	14.0%	15.0%	5.8%
	teens		8.4%	12.2%	12.2%	3.8%
10-12	*kids*	7.6%	8.2%	12.1%	12.8%	4.6%
	teens		7.5%	8.0%	8.9%	1.4%
13-15	*kids*	7.9%	1.7%	7.0%	7.4%	5.7%
	teens		5.0%	10.2%	11.3%	6.3%
adults	*kids*	7.9%	3.2%	2.8%	3.5%	0.3%
	teens		6.5%	7.6%	7.8%	1.3%
			Top 50			
8-9	*kids*	7.6%	15.3%	21.0%	21.5%	6.2%
	teens		17.1%	22.1%	23.2%	6.1%
10-12	*kids*	7.6%	13.1%	16.3%	17.1%	4.0%
	teens		15.7%	20.6%	21.2%	5.5%
13-15	*kids*	7.9%	4.8%	13.1%	13.5%	8.7%
	teens		15.2%	17.4%	18.5%	3.3%
adults	*kids*	7.9%	4.3%	4.9%	5.6%	1.3%
	teens		7.1%	8.1%	8.9%	1.8%

Table 5.6: Recall comparison using the Yahoo! Search logs.

performance of Bing is comparable across all the age groups. This trend suggests that queries from the youngest users are more frequently on the *long-tail* than queries from adults and teenagers. This observation emphasizes the usefulness of our method to address this type of queries since the best performing scores of our method are obtained at top 5 for the youngest group of users.

As it was the case with the AOL data, the best performing model was aligned with its targeting age group. That is, the model *kids* which targets users under 12 years outperforms the *teens* model for queries of users from 8 to 12 years. Similarly, this model provided a better ranking quality, as it is shown in Table 5.7. We believe that this result is valuable because it suggests that the method proposed can be exploited on different information domains or on a different niche of users by carefully choosing the set of seed urls (i.e. model).

Tables 5.8 and 5.9 show the results obtained with the Yahoo! Search logs using

query set	model	Bing	rw-b	rw-k-f	rw-k-b	gain
			Top 5			
8-9	*kids*	0.023	0.043	0.117	0.121	0.078
	teens		0.032	0.083	0.084	0.052
10-12	*kids*	0.039	0.044	0.105	0.110	0.066
	teens		0.037	0.052	0.053	0.016
13-15	*kids*	0.041	0.017	0.026	0.028	0.011
	teens		0.034	0.079	0.082	0.048
adults	*kids*	0.055	0.000	0.018	0.020	0.020
	teens		0.041	0.049	0.051	0.010
			Top 10			
8-9	*kids*	0.023	0.060	0.130	0.132	0.072
	teens		0.058	0.129	0.133	0.075
10-12	*kids*	0.040	0.061	0.119	0.121	0.060
	teens		0.051	0.066	0.069	0.018
13-15	*kids*	0.041	0.019	0.038	0.040	0.021
	teens		0.041	0.086	0.088	0.047
adults	*kids*	0.055	0.010	0.033	0.035	0.025
	teens		0.051	0.054	0.056	0.005
			Top 50			
8-9	*kids*	0.023	0.079	0.149	0.152	0.073
	teens		0.076	0.137	0.137	0.061
10-12	*kids*	0.040	0.079	0.135	0.137	0.058
	teens		0.080	0.106	0.109	0.029
13-15	*kids*	0.041	0.022	0.068	0.070	0.048
	teens		0.065	0.118	0.120	0.055
adults	*kids*	0.055	0.005	0.049	0.054	0.049
	teens		0.005	0.059	0.060	0.055

Table 5.7: NDCG comparison using the Yahoo! Search logs.

only the query reformulations that lead to *long* clicks. As we observed with the AOL logs, the performance values in terms of recall and NDCG were lower that the results obtained with the first set. Importantly, we also observed that the ratio gains between our method and the two baselines were also higher for these query reformulations, which suggests that the query suggestions provided by our method are more likely to lead to relevant content. For instance the performance gain in respect to Bing and *rw-b* for the queries of users from 10-12 years at top 10 was 7.9% and 6.3% respectively while the gain in the set of all clicks (not only *long* clicks) was of 7.4% and 5.8% respectively.

We apply the method described in the previous section to verify the statistical significance of the results. Similarly, we found all the results (when comparing all the combination of methods and within the models) were statistical significant at the 0.01 level. However, we found that for few cases the difference between our two random

query set	model	Bing	rw-b	rw-k-f	rw-k-b	gain
			Top 5			
8-9	*kids*	2.7%	4.7%	11.0%	11.7%	7.0%
	teens		4.2%	7.8%	7.9%	3.8%
10-12	*kids*	3.2%	4.1%	8.2%	8.8%	4.7%
	teens		4.1%	5.2%	5.6%	1.5%
13-15	*kids*	3.5%	0.8%	3.0%	4.3%	3.5%
	teens		4.7%	8.5%	9.3%	4.6%
adults	*kids*	5.7%	0.1%	0.9%	2.1%	2.0%
	teens		0.2%	4.8%	5.7%	5.6%
			Top 10			
8-9	*kids*	7.1%	8.7%	13.9%	15.0%	6.3%
	teens		8.3%	11.9%	12.0%	3.7%
10-12	*kids*	7.3%	7.6%	8.9%	9.9%	2.3%
	teens		7.0%	7.9%	8.9%	1.9%
13-15	*kids*	7.1%	1.3%	6.8%	7.3%	6.0%
	teens		5.0%	10.1%	11.2%	6.2%
adults	*kids*	7.2%	2.9%	2.6%	3.4%	0.5%
	teens		6.2%	7.3%	7.6%	1.5%
			Top 50			
8-9	*kids*	6.8%	14.9%	20.9%	21.5%	6.6%
	teens		17.1%	21.9%	23.2%	6.1%
10-12	*kids*	7.2%	12.7%	16.1%	17.1%	4.4%
	teens		15.4%	20.4%	21.1%	5.7%
13-15	*kids*	7.4%	4.6%	13.1%	13.5%	8.9%
	teens		14.9%	17.3%	18.3%	3.5%
adults	*kids*	7.2%	4.0%	4.8%	5.5%	1.5%
	teens		7.0%	8.0%	8.8%	1.8%

Table 5.8: Recall comparison using the Yahoo! Search logs (long clicks).

models were not statistical significant. These cases are shown with underlined values in Tables 5.6 and 5.7.

It is important to mention that the low values of NDCG reported are due to the sparsity of the data. On average we collected 1.6 query suggestions per query. Nonetheless, numbers on the same order have been reported on query recommendation studies for long-tailed queries [Szpektor et al., 2011].

5.7 Learning to Rank Tags

State of the art search engines employ large number of features to rank web results. In the previous section we showed that our method outperforms traditional random walks and a state-of-the art search engine in the problem of recommending queries to young users. We consider that the results can be improved further using a learning

query set	model	Bing	rw-b	rw-k-f	rw-k-b	gain
			Top 5			
8-9	kids	0.019	0.039	0.116	0.120	0.081
	teens		0.030	0.081	0.083	0.053
10-12	kids	0.033	0.042	0.105	0.109	0.067
	teens		0.035	0.049	0.051	0.017
13-15	kids	0.026	0.013	0.024	0.026	0.013
	teens		0.029	0.069	0.079	0.050
adults	kids	0.042	0.001	0.016	0.020	0.019
	teens		0.038	0.047	0.051	0.012
			Top 10			
8-9	kids	0.020	0.059	0.128	0.130	0.071
	teens		0.053	0.130	0.131	0.078
10-12	kids	0.034	0.057	0.115	0.120	0.063
	teens		0.051	0.067	0.067	0.016
13-15	kids	0.039	0.015	0.037	0.086	0.071
	teens		0.042	0.083	0.086	0.044
adults	kids	0.051	0.002	0.031	0.035	0.033
	teens		0.040	0.050	0.053	0.013
			Top 50			
8-9	kids	0.022	0.076	0.150	0.152	0.076
	teens		0.076	0.135	0.135	0.060
10-12	kids	0.038	0.077	0.134	0.138	0.061
	teens		0.077	0.105	0.108	0.031
13-15	kids	0.040	0.021	0.064	0.068	0.048
	teens		0.064	0.118	0.121	0.057
adults	kids	0.054	0.007	0.047	0.051	0.044
	teens		0.047	0.054	0.060	0.012

Table 5.9: NDCG comparison using the Yahoo! Search logs (long clicks).

to rank framework in which our random walk method would represent one of the features. We envisage three types of features to improve the system: language models, topic features, and distance to seed tags. In the following paragraphs we describe the features introduced for each one of these categories.

5.7.1 Language Model Features

We expect query suggestions to appear more frequently in the same documents in which the query occurs, particularly on documents written with a vocabulary appropriate for children. We build a language model using the websites listed in the Dmoz *kids and teens* directory to estimate the likelihood of a query suggestion candidate to co-occur in the neighborhood of the user's query. The language model is defined

in the following way:

$$p(q|Mc_t) = \prod_{i=1}^{|q|} p(q_i|Mc_t) \tag{5.19}$$

where Mc_t is the language model of the context in which the query suggestion t occurs. The context is constructed using n-grams in a window of n words before and after q_i. Note that this window can be set as the whole document or as fine-grained section of the document. The probability is estimated in the following manner:

$$p(q_i|Mc_t) = \frac{cf_{Mc}(q_i, t) + \mu\, p(q_i|M)}{|C| + \mu} \tag{5.20}$$

where M is the entire Dmoz collection, $|C|$ is the total of n-grams in the context (i.e. pseudo document) we are considering and μ is the Dirichlet smoothing parameter. We estimated these parameters for optimal performance using a set of the queries of the AOL query log. The context window was set to 20 words before and after t and μ to 800.

We also expect suggestions more frequently used by children to appear more frequently in the Dmoz *kids and teens* collection. For this reason we consider as a feature the probability of the query suggestion in this collection: $p(t|M)$.

5.7.2 String features

We employed two simple string features: *(i)* query length and *(ii)* query suggestion length. We believe children favor shorter query suggestions since these suggestions tend to represent simpler words. We represent the length of the query by the number of tokens and by the number of alpha numerical characters. We report only the later since the results obtained by both approaches were identical.

5.7.3 Topic Features

We hypothesize that the rank of the suggestions can be improved informing the system about the topics that the query targets and the candidate suggestions that best represent the content of these topics. Consider the query *penguin*. The suggestions *games*, *online* and *cheats* are the 3 top ranked results provided by our random walk. Although, these suggestions are coherent and appropriate for children, the user submitting this query may be targeting general information about the flightless bird (e.g. for a school homework), or the user may simply be looking for pictures or videos of penguins instead of gaming related content. In the Dmoz kids & teens directory, penguin is associated with the topics *School Time\Science* and to *Games\Computers and Videos*. Using this information our system can boost the suggestions related to

the topic *school time*, which are under-represented given the dominance of the gaming aspect for this query in our data.

The strategy is to generate a topic representation of the query and the candidate suggestion to estimate the topic coherence between the two elements. For this purpose, a query topic classifier was implemented by indexing the documents in the Dmoz *Kids and Teens* directory, which were appropriate for children up to 12 years old. We indexed around 15K websites with this approach. In this collection each document is located under one or more categories. Documents were mapped to topics by utilizing the most popular (in terms of size) and specific category to which the document belongs. We trim the depth of the category to two levels as an attempt to avoid data sparsity. This procedure led to 60 different categories or topics.

The classification of queries and query suggestions is carried out based on the top 100 matching documents (using standard *tf-idf* ranking). Concretely the topics associated to each one of the results returned for the query are fetched and the retrieval score are aggregated on a per topic basis. In this manner we obtain a weight of the importance of each topic for the query. A vector of topic features is constructed by normalizing the scores in the vector between 0 and 1 (non-matching topics are assigned an score of zero). Formally, each topic feature is defined in the following manner:

$$cat_q(z) = \sum_{j=1}^{|R_{100}|} I(z = category(doc_j)) * score(doc_j) \tag{5.21}$$

where I is the identity function and $category(d)$ is the function that maps a document to a Dmoz category. The $cat_q(z)$ values are normalized by dividing over the sum of all the topic features calculated given a query q.

Using these features, each document is represented as the vector:
$V_Q = (cat_q(1), cat_q(2), .., cat_q(|z|))$ where $|z|$ is the total number of topics considered in the system.

Concretely we employed the topic vectors as features in two ways:

- the vector representation of the query as a set of features of size $|z|$
- by computing the cosine similarity between the topic vector of the query and the candidate query suggestion: $sim = \frac{\sum_{i=1}^{|z|} A_i \times B_i}{\sqrt{\sum_{i=1}^{|z|} A_i \sum_{i=1}^{|z|} B_i}}$

The motivation of the latter is to capture the topical cohesion between the query and the query suggestion.

5.7.4 Similarity to Seed Keywords

Seed tags as *for kids* or *kids* are often employed to signal that a web resource is designed for children. For this reason we expect children-oriented suggestions to appear

more frequently near these seed tags. We estimate the distance between the candidate query suggestion and a pre-defined set of seeds tags (concretely, *for kids*, *for children*, *kids*, *4kids*). We expect this feature to stress candidate query suggestions that are more children friendly. The estimation was carried out by summing the probabilities of the query suggestion being used to describe a web resource in conjunction with the seed tags:

$$sim(t_j, s) = \frac{\sum_{i=1}^{|s|} p(t_j|s_i)p(s_i)}{\sum_{j=1}^{|T|}\sum_{i=1}^{|s|} p(t_j|s_i)p(s_i)} \tag{5.22}$$

$$p(t_j|s_i) = \frac{cf(t_j, s_i)}{cf(s_i)} \tag{5.23}$$

where s is the set of seed tags and $|T|$ is the set of tags in the vocabulary. The probability $p(t_j|s_i)$ is estimated on the delicious corpus which was also utilized for the random walk.

5.7.5 Learning to Rank Evaluation

We evaluate the features described in the previous section using the query set of users from 10-12 and 13-15 years old. Concretely, we employed a subset of the query reformulations utilized in the evaluation of the random walk. The training (and testing) data is in the order of the tens of thousands of query reformulations. The training data was constructed by extracting those queries for which there is at least one correct result in the list suggestions provided by the random walk, considering the top 50 ranked results. We use the gradient boosted regression tree learner to train the model and we perform 10-fold cross validation on the same data. Default parameters were used[1].

The best performance is obtained when all the features are combined: the NDCG is increased from 0.564 to 0.670, 0.541 to 0.642 and 0.523 to 0.623 for children aged 8 to 9, 10 to 12 and for teenagers respectively on the training data using 10-fold cross validation. The performance of the entire dataset (recall that the training data is a subset of the query reformulations extracted for users aged 10 to 12 years old) is increased from 0.132 to 0.189, 0.121 to 0.172 and 0.082 to 0.124 for users aged 8 to 9, 10 to 12 and 13-15 respectively.

Table 5.10 presents the NDCG scores obtained when each feature is employed independently. We found that the random walk score is by a large margin the best predictor (*e.g.* 0.541 vs. 0.313 of the next best performing feature in the case of users aged 10 to 12). The topic similarity metric and the language model trained on the Dmoz corpus were the next best performing features. For instance for users aged 10

[1] `http://people.cs.umass.edu/~vdang/ranklib.html`

feature	age	ndcg	ndcg-global
all	8-9	0.670	0.189
	10-12	0.642	0.172
	13-15	0.623	0.124
topic vector	8-9	0.137	0.230
	10-12	0.090	0.021
	13-15	0.111	0.022
topic sim	8-9	0.146	0.032
	10-12	0.123	0.028
	13-15	0.142	0.028
$p(q\|Mc_t)$	8-9	0.343	0.090
	10-12	0.313	0.070
	13-15	0.251	0.050
$p(t\|M)$	8-9	0.031	0.012
	10-12	0.052	0.014
	13-15	0.021	0.004
$sim(t\|S)$	8-9	0.052	0.016
	10-12	0.063	0.017
	13-15	0.051	0.015
q. length	8-9	0.006	0.002
	10-12	0.005	0.002
	13-15	0.003	0.001
s. length	8-9	0.001	0.000
	10-12	0.004	0.002
	13-15	0.003	0.001
rw-kl-b	8-9	0.564	0.132
	10-12	0.541	0.121
	13-15	0.523	0.082

Table 5.10: NDCG scores for each feature. Underlined values were not proven statistical significant at $p < 0.01$.

to 12 these features represent a gain of 0.123 and 0.313 NDCG respectively. A similar outcome was observed for the other age groups.

The other features perform poorly when they are use in isolation in the system. This result can be explained by the fact that these features do not model the relation between the query and the query suggestion. For instance the feature expressed in equation 5.22 only provides information about the similarity of the query suggestion to a predefined set of seed tags. Nonetheless these features are beneficial when they are used in conjunction with the random walk score.

Table 5.11 presents the NDCG values obtained by the system when dropping each one of the features. Consistently with the results reported in Table 5.10, leaving out the random walk feature leads to a performance loss of 74.3% for children between 8 to 9 years old, 76.2% for users aged 10 to 12 and 74.3% for teenagers. Interestingly,

feature	age	ndcg	ndcg total	%diff
all	8-9	0.670	0.189	0.0%
	10-12	0.642	0.172	0.0%
	13-15	0.623	0.124	0.0%
no topic vector	8-9	0.642	0.181	-4.2%
	10-12	0.639	0.178	-0.5%
	13-15	0.602	0.143	-3.4%
no topic sim.	8-9	0.564	0.159	-15.8%
	10-12	0.550	0.153	-14.8%
	13-15	0.525	0.138	-15.7%
no $p(q\|Mc_t)$	8-9	0.555	0.162	-17.2%
	10-12	0.568	0.158	-11.5%
	13-15	0.581	0.138	-6.7%
no $p(t\|M)$	8-9	0.656	0.182	-2.1%
	10-12	0.612	0.162	-4.7%
	13-15	0.601	0.143	-3.5%
no $sim(t\|S)$	8-9	0.582	0.164	-13.1%
	10-12	0.572	0.159	-10.9%
	13-15	0.590	0.143	-5.3%
no q. length	8-9	0.670	0.189	0.0%
	10-12	0.641	0.171	-0.2%
	13-15	0.621	0.123	-0.3%
no s. length	8-9	0.670	0.189	0.0%
	10-12	0.640	0.171	-0.3%
	13-15	0.621	0.123	-0.3%
no rw-kl-b	8-9	0.172	0.069	-74.3%
	10-12	0.153	0.043	-76.2%
	13-15	0.160	0.038	-74.3%

Table 5.11: Leave-one-out NDCG scores. Underlined values were not proven statistical significant at $p < 0.01$.

the biggest performance loss after the random walk feature is obtained by dropping the topic similarity feature (*e.g.* 14.8% for the 10 to 12 group), which shows the importance of *informing* the system about the topical relation between the topics and the query. We observed that the topic representation of the query did not lead to significant improvements. This may be due to the large number of features that this vector represents (up to 80). The $sim(t|S)$ feature also leads to a significant loss of performance despite its simplicity (*e.g.* 13.1% for the 8 to 9 group). Given this results we believe it is worth to study in future work a more formal procedure to select the set of tags since the performance gain is already significant with a small (although representative) set of seed tags.

As it was the case with the results in Table 5.10, we observed that the string features did not influence the overall performance of the classifier and consequently

these features did not provide any performance gain in our experiments. The results presented in Table 5.10 and 5.11 were proven statistical significant using a t-test at $p < 0.01$ except for the values underlined in the tables.

5.8 Conclusions and future work

In this chapter we showed how tags from social bookmarking system can be exploited to produce query suggestions for a specialized group of users using a set of seed web resources and a biased random walk based on the point-wise KL divergence metric between a foreground model and background model. We further improved the ranking of our results using a learning to rank approach to utilize the random walk score along intuitive features to boost query suggestions oriented on children topics.

In regard to *RQ-5.1*, we showed that our method can be used to improve current search assistance functionality for children since it performs the best for the youngest groups of users (8 to 9 years old). This segment of users is not served as well as older users, partly due to the long tail problem, which is more pronounced for these users according to the test collection created. We observed that two of the biased random walks tested (*seedRank* and *spamRank*) also outperforms state-of-the-art query suggestions, however the method proposed outperforms largely these two biased random walks as well. We also observed that predicting query reformulations landing on long clicks represents a harder problem since the absolute performance scores (recall and *NDCG*) were lower on this restricted test set. However, we show that the performance gain of our method tend to be even higher under this evaluation setting.

In regard to *RQ-5.2*, we found that a learning to rank approach is a suitable framework to combine the score of the random walk with additional features, since their combination lead to significant improvements in terms of *NDCG*. It is worth to mention that topical features derived from the *Dmoz Kids and Teens* were proven highly beneficial while other simple features based on the characteristics of the query (*e.g.* query length) were not found useful for this problem.

In respect to *RQ-5.3*, we obtained consistent results in the AOL logs and the Yahoo! logs. This suggests that a similar method to extract log data using quality seed web resources can be exploited to study and evaluate query reformulation for specialized niche of users if the seed set of urls and tags are carefully selected.

5.8.1 Future work

For future work it is worth to apply the method proposed in different information domains. In this work we focused on content for young users but potentially we could apply the same principles in domains such as query suggestions for professionals in the business domain or any other specific fields of expertise.

Similarly, research can be carried out to expand the methods we have presented in this chapter. For instance, features based on the complexity of the language and the readability of the pages returned by the query suggestions can be explored as an attempt to improve the quality of the suggestions. Similarly, behavioral features derived from the search sessions can also be beneficial.

Another line of research can involve studying the impact of having not only textural query suggestions but also multi-modal query suggestions. For certain information tasks it may more useful for young users to suggest videos, images or audio content. However, we believe that the methods proposed represent an important first step towards helping children to build queries in a real world search engine.

In the next chapter we will explore the problem of selecting the right set of search engine given the query submitted by a young user. This approach represents a step forward helping children to find the right content on the Internet.

Chapter 6

Vertical Selection for Young Users

This chapter is based on Duarte Torres et al. [2013]

6.1 Introduction

In chapters 2 and 3 we observed that young users need a big effort to access high quality suitable for them, and that this group of users often struggle during the search process. Moreover, in Chapter 4 was observed that young users have a higher likelihood of submitting queries to multimedia search services after carrying out browsing activities. Given these observations, young users would highly benefit from a better integration of results that are suitable for them, particularly from an integration of content of different genres that support the exploration of information. As it was mentioned in Chapter 1, aggregated search represents an adequate paradigm to carry out this integration.

Recall that in aggregated search, content is retrieved from different search services and the content retrieved is integrated in a meaningful and consistent way. These search services, or verticals, are defined as domain specific collections, (*e.g.* entertainment, shopping, news) or collections from specialized types or genres (*e.g.* videos, images, songs). Generally, aggregated search systems are assumed to have complete access to the verticals, which implies having access to the query logs and to detailed statistical descriptors of the verticals.

In this thesis, we envisage a search system that integrates heterogeneous content from verticals, which are not fully accessible to the system (third party verticals). The system envisaged integrates content of different genres as it is done in aggregated search. In this chapter we will focus on verticals that contain high quality information for children from 8 to 12 years old. In this system parents, teachers and other specialist on children care are allowed to add resources for this group of users. For instance, they could add a vertical dedicated to coloring pages: `http://ivyjoy.com/colouring/search.html`, which only returns sheets of paper to be colored and that

are suitable for children, or a vertical dedicated to search for videos: `http://www.youtube.com`, in this case the vertical provide content for all kind of public segments. We believe that an aggregated search system is a better solution for this niche of users than simply crawling and indexing websites, or listing suitable content (*e.g.* `http://www.kids.yahoo.com`) because *(i)* it is more scalable, *(ii)* we can leverage and exploit the knowledge of parents and other experts by using hundreds of services suggested by them.

Under this scenario, once a query is submitted, the system has to decide which are the most relevant verticals for the query. This problem has been characterized as a multi-class classification problem [Arguello et al., 2010; Zhou et al., 2011], in which the objective is to predict the set of relevant verticals (or single vertical in [Arguello et al., 2009]) from a set of predefined verticals that are accessible by the system. This problem is referred as vertical selection and it has been widely studied [Arguello et al., 2009; ilad Shokouhi and Si, 2011; Zhou et al., 2011].

In this first part of this chapter, we explore two methods to address this problem in the information domain of users aged 8 to 12 years old. Our problem differs from those addressed in previous studies in that: *(i)* the vertical selection is carried out under the restriction of targeting a specific information domain (*e.g.* content for children); *(ii)* these users search for domain specific content in a set of verticals that may not be completely suitable for them, thus some verticals may provide only suitable content (*e.g.* coloring pages) but others may or may not contain suitable content for their information needs (*e.g.* YouTube) since it contains content for all type of public and *(iii)* a test collection for this domain has not been built until this work and the process of gathering assessments is not straight forward given the nature of the targeted users.

The two methods proposed are evaluated with a novel test collection for the problem of vertical selection in the domain of content for children. We describe its construction in sections 6.3 and 6.4. Given the novelty of the collection we describe two methodologies to gather relevant assessments using crowd-sourcing: first assuming that the web vertical is always displayed (which is the case in state-of-the-art aggregated search interfaces) [Arguello et al., 2009] and the second methodology easing this restriction.

In the first method, we estimate the children suitability of each vertical by measuring the amount of child-friendly content and the amount of content available for the average web user. A simple approach is proposed to combine both estimations (children and non children vertical sizes) and we show that their used in *ReDDe* [Si and Callan, 2003], a state-of-the-art vertical selection method, lead to significant improvements:

R.Q-6.1: To what extent can we improve state-of-the-art techniques of vertical

selection through the estimation of the content available in the verticals for users between 8 to 12 years old?

In the second method, we present a novel method of vertical selection for domain specific scenarios using tags from social media and language models. Concretely, verticals are represented based on tags describing the urls from a sample of each vertical and a language model on these tags is employed to rank the relevancy of the verticals given a query. As far as we know social media has not been employed before as a source of evidence in the problem of vertical selection:

R.Q-6.2: What is the benefit of using tags from social media to represent the query and the verticals in the problem of vertical selection?

6.1.1 Validating the benefits of aggregated results with children users

The test collection employed to address *R.Q-6.1* and *R.Q-6.2* relies on the assumption that adult users are able to identify accurately content that is suitable for children. Nonetheless, it is still unclear the benefits that aggregated pages (built from the verticals our collection) provide to children, in terms of interaction and exploration of results. Similarly, although it is reasonable to assume that adults are able to discern between appropriate and non-appropriate content for children, we have not explored how these judgments can differ from the judgment of actual children in the target age range. These two concerns are addressed in the following research questions:

R.Q-6.3: Do users aged 8 to 12 years old explore more pages with blended results from the verticals of our collection than pages retrieved from a state-of-the-art search engine? In which type of result pages do they agree more in terms of the content clicked?

R.Q-6.4: Which verticals are preferred by children aged 8 to 12 years old given an heterogeneous set of topics, and how do these vertical preferences differ from the preferences of adult users?

To address these research questions, we present in Section 6.9 a game designed to measure the interaction and exploration of users, aged 8 to 12 years old, in three types of page results: *(*i) plain web results from the Google search engine,*(ii)* state of the art aggregated results from Google, and *(3)* aggregated results using the verticals from our test collection and the relevant judgments gathered using *CrowdFlower*. The game asks users to collect snippets from one of the three types of pages given an open information task. For each information task, users are shown, randomly, one of the three types of pages under evaluation. Users are awarded points for each result

that is clicked based on user agreement. Thus, the greater the number of users that click on a given result, the more points this result awards.

The design of the game is deeply inspired by the guidelines suggested by von Ahn and Dabbish [2008] to design games with a purpose (GWAPs), which are meant to solve collectively complex computational problems through online games in an entertainment way. In our case, the problem is to select the right set of verticals for open information requests. In Section 6.9 is described how these guidelines applied to our case.

Our first aim is to compare the interaction of users in the three types of pages in terms of content exploration and agreement, as it is expressed in *R.Q-6.3*. The central hypothesis is that the aggregated pages built from the vertical results of our collection promote the exploration of results. We also expect users to agree more on the content that is relevant in the aggregated pages, since the results embedded in this type of pages were chosen carefully from a set of verticals providing appropriate content for young users. The exploration of results in the three types of pages is compared using the following metrics: rank distribution, game session length, average time spend per click and average time spend per game round. The rank distribution shows how often users click on lower ranked results. The game session length refers to the number of clicks registered in each game round. For this study, a game round correspond to the presentation of page result (either page type) for a randomly selected topic. This metric is used to estimate the number of tasks that were skipped in each one of the pages, and in general the average number clicks registered in each page type. We expect users to have a longer session length on our aggregated results, as a result of the greater amount of quality choices to select from. The last two metrics are used to observe if children are more engaged with the results shown in the aggregated pages built with our collection by spending more time on these result pages.

Recall from Chapter 3, that young users tend to explore fewer results and to have shorter sessions in terms of duration and number of clicks, likely as a consequence of the inadequacy of the results presented to them. We expect to reduce this behavior by providing more suitable results that are moderated and more of the like of young users. Additionally, we estimate the inter-assessor agreement to discern in which type of pages users agree more on the elements clicked and the verticals selected. The differences in the agreement values indicate that children are better finding relevant and suitable information in specific page types.

We address *R.Q-6.4* by comparing the vertical distribution clicked by a group of children and adults in the game. The group of children consisted of 11 students aged 9 to 10 from an international primary school in the Netherlands. Crowd-sourcing was employed to represent the group of adults. Additionally, we measure the vertical agreement, on a per topic basis, between the two groups of users. A detail analysis was carried out to identify the verticals in which children and adults disagreed.

6.1.2 Chapter outline

This chapter is organized as follows: Section 6.2 presents the most relevant related work to the methods proposed in this chapter. Section 6.3 describes the selection of queries and verticals in the construction of the collection for the evaluation. Section 6.4 describes the characteristics of the data collected. Section 6.5 presents an analysis and comparison of the relevant assessments gathered using the two approaches mentioned above: assuming that the web verticals is always displayed and when this is not the case. Section 6.6 describes the approach adopted to estimate the amount of content aimed at children in each one of the verticals in the collection. Section 6.7 shows how these size estimations can be applied to state-of-the-art vertical selection methods and how to employ tags from social media for this problem and information domain (*RQ-6.1* and *RQ-6.2*). Section 6.8 describes the experiments and discusses the results.

The description of the game designed to evaluate the benefit of aggregated pages is found in Section 6.9. The settings of the case study and the description of the group of users are described in Section 6.10. The interaction results are explored in Section 6.11 and the results of the follow-up survey delivered to the group of children are found in Section 6.12. In Section 6.13 is discussed the implication of the results of the case study in respect to *R.Q.-6.3* and *R.Q.-6.4*. Finally, Section 6.14 summarizes this chapter and provides points for future research.

6.2 Related Work

6.2.1 Vertical Selection in IR

Arguello et al. [2009, 2010] proposed a machine learning approach to combine the scores of several resource selection methods. *ReDDe*, which was initially proposed for federated search and databases [Si and Callan, 2003], is adapted for large-scale aggregated search systems. Our work differs from theirs in that: *(i)* we are interested on domain specific areas (*e.g.* content for children); *(ii)* our methods are applied on public available resources instead of proprietary environments (absent of query logs); *(ii)* we relax the restriction of having only one relevant vertical per query since we observed that even though the number of relevant verticals is small, the average is far greater than 1 (around 3.5); *(iv)* a collection is provided to the research community. We hope this will motivate the further study of vertical selection in the domain of children.

Arguello et al. [2009] showed a methodology to gather vertical relevance assessments by asking trained annotators to find the most suitable vertical for the query. Arguello et al. [2011] employed paired comparisons to evaluate aggregated search in-

terfaces. Thomas and Hawking [2006] used this methodology in IR initially. In this work, we employ the paired comparison methodology to gather assessments and we show that this method may lead to preferences towards visual oriented verticals. We propose an alternative methodology that does not suffer of this bias.

6.2.2 Evaluation of aggregated result pages

Sushmita et al. [2009] compared an interface blending results from verticals against a simple tab interface with indirect links to the vertical results. A case study with sixteen participants was carried out. They found that the accessibility is improved in the aggregated interface. They observed that a larger number of results is retrieved with the aggregated interface and that the quality of the results retrieved was higher. In the same line, Sushmita et al. [2010] found that certain characteristics of click through behavior, such as the interaction on video results and the rank position of relevant results, is significant and differ when using aggregated interfaces. The setting of our problem is similar to the one studied by Sushmita et al. [2009, 2010], however, we are interested in evaluating the benefits of aggregated pages with results from our test collection for the case of users aged 8 to 12 years old. We contrast this type of pages against blended pages from a state of the art search engine and against simplified pages showing results from the web vertical. Our experimental design is different given the characteristics of the target users. We designed a game in which users solve open information tasks. Through their click behavior we compare their interaction in the three types of pages provided. The methodology presented is novel since this type of evaluation has not been addressed by using games with a purpose.

Arguello et al. [2012] explored the relation between interaction and task complexity in aggregated search results. Pages with blended results from verticals and pages providing links to the verticals are compared in a case study with 29 participants and 6 tasks of different level of complexity. They found that complex tasks led to a bigger amount of interaction, measured by results explored and number of clicks. Although no differences were found in the amount of interaction between the two types of interfaces, they observed a larger number of clicks on vertical results when using the blended interface. On overall the log metrics utilized to compare the two interfaces were not conclusive. Nonetheless, users reported a preference towards the aggregated pages. In our work, we focus on open information tasks, thus the complexity of the task is not considered in our study. The motivation of our work is to measure the adequacy of the vertical results suggested by adult users and to compare the vertical relevance judgment between children and adults. We also explore the benefits of aggregated interfaces in terms of the amount of interaction registered in the pages.

In a similar study, Arguello and Capra [2012] explored the relation between the

interaction and the coherence of the vertical results. Coherence refers to the level of agreement between the vertical results and the query aspects. For instance, if the query *cars* is submitted and users are searching for the movie *cars*, results displaying only pictures of vehicles, unrelated to the movie, would have low coherence with the query. They found that when the web vertical has a high coherence with the sense of the query that is targeted, the image results do not have a big impact in the interaction on the web vertical. However, if this is not the case, the results presented from the image vertical have a significant impact in the interaction with the web results. As we point out before, our focus is on the interaction differences between aggregated/non-aggregated pages, and on differences between children and adults. We disregarded the coherence dimension given that our evaluation utilizes open information tasks. Nonetheless, this is an important matter that needs further research.

In a recent study, Bron et al. [2013] investigated users preferences towards aggregated and non-aggregated pages during multi-session search tasks. In a longitudinal study they found that the tabbed displays are used more frequently, nonetheless the participants often switched between the two types of interfaces. Also, users were observed to use the tabbed results to *zoom in* the information provided by specific verticals and the aggregated interface was used to explore and have and overview of the information provided by the verticals.

6.3 Collection construction

The collection built consists of a carefully chosen set of queries and verticals. A set of vertical results was also retrieved for each (*query*,*vertical*) pair and relevant assessments mapping a set of relevant verticals to each query were gathered. Using this annotation schema, we can test and compare any pair of vertical selection methods. In the following sections we describe in detail the methodology carried out to build the collection and we justify our decisions in the design of the collection.

6.3.1 Query set selection

The query set was extracted from the AOL query logs [Pass et al., 2006]. We extracted queries landing on domains listed in the *Kids and Teens* directory. Given that only domains are displayed in the AOL log we carefully extracted those entries in which the exact domain is listed in the Dmoz directory (*i.e.* exact matching). An analogous procedure has been suggested for this information domain in Chapter 2. The Dmoz directory has been used in previous research in information retrieval for children [Eickhoff et al., 2011].

Query	Dmoz Category
1950's television shows	News
science fair projects ideas	School - Time/Science
barometric pressure	School - Time/Science
bingo song copyright	Arts/Music
rabbit ears	Sports - Hobbies/Crafts
secret code game	Games/Word_Play

Table 6.1: Example of queries and their Dmoz category.

We left out navigational queries by filtering out query-click pairs in which the domain is mentioned in the query. For instance, the query *sesame street* is filtered out if it lands on the domain *www.sesamestreet.org*. We also filter out query containing tokens such as *.com www.*, *http* and *.org*. The Levenshtein distance between the query and domain of the url was also employed to filter out queries that misspell a domain (*e.g.* pbkids). Concretely, the queries were selected from search sessions satisfying either of the following restrictions:

1. There is at least one click event after submitting the query that lands on a domain listed in Dmoz and the duration of the click event is of at least 60 seconds.

2. There is a click on the Dmoz domain is the last event of the session.

The first restriction is employed to capture only the cases in which click events have long duration, that is, if users spend more than 1 minute on a web result is a strong indication that the result clicked is relevant for the query [Hassan et al., 2010a]. The second condition is also employed because previous query log studies have shown that the last click of the session can be often associated to successful searches [Downey et al., 2008]. We extracted a set of 3.8K queries by using both restrictions. Table 6.1 provides examples of the queries extracted and the Dmoz *Kids and Teens* category in which they belong.

6.3.2 Selection of verticals

The list of verticals selected is shown in Table 6.2. The verticals were manually selected to cover the information needs found in the query set and the distribution of topics targeted by children between 8 to 12 years old on the web, according the topics identified in Chapter 3 We found that several genre verticals need to be split into fine-grained information services to fit the topic interests of children. For instance in our system the vertical *games* may refer to the *video games* or to the *online gaming* vertical. Similarly the genre *images* may refer to *coloring pages*, *printable worksheets* or the standard *pictures* vertical. In previous literature, these services are wrapped

under a single vertical (either images or games respectively) [Arguello et al., 2010]. This fine-representation of verticals is highly convenient for young users because it increases the accessibility to rich media. Recall that this niche of users has been found to struggle to identify and to search for information on non-web verticals [Foss et al., 2012]. Dedicated verticals offering content for children may be found on the Internet (*e.g.* video-games, stories, health), however some of the verticals displayed in Table 6.2 were constructed artificially by modifying the search parameters of large search services. For instance, the vertical *coloring pages* is constructed by using the *line-drawing* parameter of the Google Image search service. For the case of the *printable worksheets*, we employed the *line-drawing* parameter of the Google Image service and we modified the user's query by expanding it with the terms *worksheets*.

It is important to clarify, that some of the results that are considered as web vertical results in the literature are considered from specialized verticals in this work. For instance, the results from the vertical *health* (*kidshealth.org*) can be seen as a web vertical result according the definitions of other studies. However in this work, it is considered as a result from the *health* vertical since it represents a result from a specialized service on health content for children, regardless of the fact that it can be found on the web vertical of state-of-the-art search engines.

6.4 Data characteristics

We describe the collection based on the number of queries covered per vertical and the distribution of the numbers of verticals covering a query. Figure 6.1 depicts the vertical coverage, which refers to the proportion of queries covered by each vertical. A vertical is said to cover a query if it returns at least one result when the query is submitted. From this figure, we observe that large verticals such as *web* and *videos* tend to cover most of the queries in the dataset. Note that even relatively small verticals such as *how-to* or *stories* have high query coverage, which suggest that the verticals chosen are appropriate for the information needs targeted by the chosen queries. Nonetheless, we observed that the verticals *movies* and *reference*, which are widely used in previous vertical selection studies, cover less than 25% of the queries in our data set. This result may indicate that these verticals are less suitable for the audience we are interested in. Figure 6.2 shows the number of queries in the collection that are covered by a specific number of verticals. For instance from this figure we can observe that there are no queries in the collection that are covered by exactly two or three distinct verticals. Similarly around 75 queries are covered by exactly 10 verticals. Interestingly we observed almost a normal distribution in which most of the queries are covered between 10 and 21 verticals (84% of the queries). On average, queries from the entire collection are covered by 16.8 verticals. This result shows that the problem of vertical selection in the domain of children topics is not

Vertical	Websites
web	google.com
games online	onlineflashgames.com
games	gamespot.com
images	google.com/imghp
coloring pages	google.com/imghp (line drawing option)
worksheets	google.com/imghp (query + worksheets)
books	books.google.com
question/answers	worldoftales.com
stories	worldoftales.com
shopping	amazon/toys
music	allmusic.com
videos	youtube.com
movies	rottentomatoes.com
enciclopedia	wikipedia.com
reference	dictionary.kids.net.au
how-to	www.instructables.org
school aid	livescience.com
	howstuffworks.com
	dsc.discovery.com
school activities	sciencekids.co.nz
	enchantedlearning.com
	howtosmile.org
lyrics	lyricsdrive.com
health	kidshealth.org

Table 6.2: List of verticals and their urls.

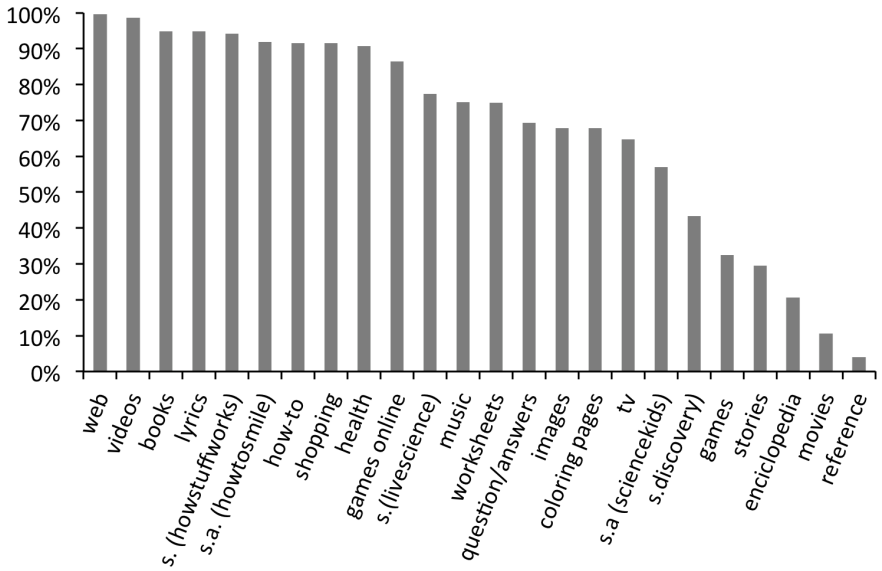


Figure 6.1: Queries covered by each vertical.

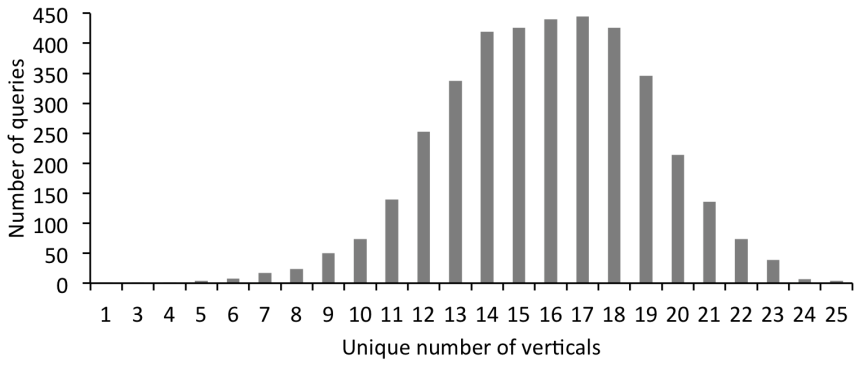


Figure 6.2: Distribution of unique verticals per query.

straight forward since each query has on overage a set of 16 verticals from which to choose relevant verticals.

6.5 Gathering vertical relevance assessments

We gather assessments by employing the crowd-sourcing engine Crowdflower[1] and a sample of 90 queries from the set of queries described previously. This sample was chosen to be representative of all the possible topics in the collection (based on the Dmoz categories of the queries). We carried out two experimental protocols for gathering the assessments. In the first protocol assessors were asked to chose between two set of vertical results: the results of the target vertical against the results returned by the web vertical (*i.e.* Google Web), each result set was displayed in a column next to each other in the survey page. Concretely, the top 4 results of each vertical were shown in each column and the order in which the columns appear was randomized. Nonetheless, the ranking of the results of each vertical was preserved. Adult assessors were able to choose *the most relevant set of results for the query (given that the content has to be suitable for users between 8 to 12)* between the two columns. The special option *none of the sets are relevant and suitable for children* was also given. This option was provided to avoid false positives since with this option users are not forced to chose a vertical when both sets are inadequate. The motivation of comparing each vertical against the web vertical is that in modern search engines the web results are always displayed, and the results from other verticals are only displayed when the vertical results *add value* to the current web results, that is, when their results are preferred over the standard web results [Arguello et al., 2011]. The main drawback of this protocol is that we are unable to identify the cases in which the results from the web verticals are unsuitable for the query.

[1] http://crowdflower.com/

In the second protocol, we asked assessors to judge vertical results independently. This protocol is motivated by the fact that in an information system for children we may not always want to present the results from the web vertical, since the user may be requiring a different type of content (*e.g.* coloring pages, videos) or because the results from the web vertical are not suitable for children given the query. For this reason we asked users to assess each set of vertical results independently using a scale-graded system of 5 points, from *bad* to *excellent* in terms of relevance to the query and appropriateness for children in the targeted age range. An advantage of this method arises in the possibility of evaluating directly the quality of the web vertical when the information needs are targeted at young users. We can also rank the verticals based on the graded score assigned.

For both protocols, adult users were asked to make the judgments. We consider reasonable to assume that adults can easily discern between content oriented for adults and children and thus that they are able to judge the results in the context of the domain. We also provide the queries to the assessors without a close description. The motivation for this was to let the assessors identify all the possible content that can be relevant for children given a query. To ensure the quality of the assessments, 120 golden judgments were created for each experimental protocol. These gold judgments were employed to avoid spammers by: *(i)* forcing the users of *CrowdFlower* to complete a training session in which they are shown only *units* (survey page) from the gold judgment set. They are allowed to start the task if they answer correctly at least 6 of these units; *(ii)* during the task the gold units are mixed with the units under evaluation, users answering incorrectly more than 8% of the gold units shown to them during the survey are ignored. Only users from the United States were allowed to carry out the survey to ensure language proficiency and domain knowledge (the queries were extracted from the US market). Each work of unit was paid with 0.01 dollar cents and each unit was evaluated by at least 3 assessors. In the following paragraphs, we describe the assessments gathered and we compare the results obtained by both experimental protocols.

6.5.1 Distribution of relevant verticals

We created a gold test set by mapping each query in the sample to a set of relevant verticals by using the assessments collected in *Crowdflower*. For the first protocol (*i.e.* paired assessments), we map a query to a vertical if at least a certain percentage of annotators select the vertical as relevant for the query. The threshold 60% and 80% were used in the results reported. As a point of reference 60% means that more than half of the assessors agreed on classifying the target vertical as relevant, and 80% means that most of the assessors agreed.

For all the thresholds we observed in Figure 6.3 a long tail distribution in which

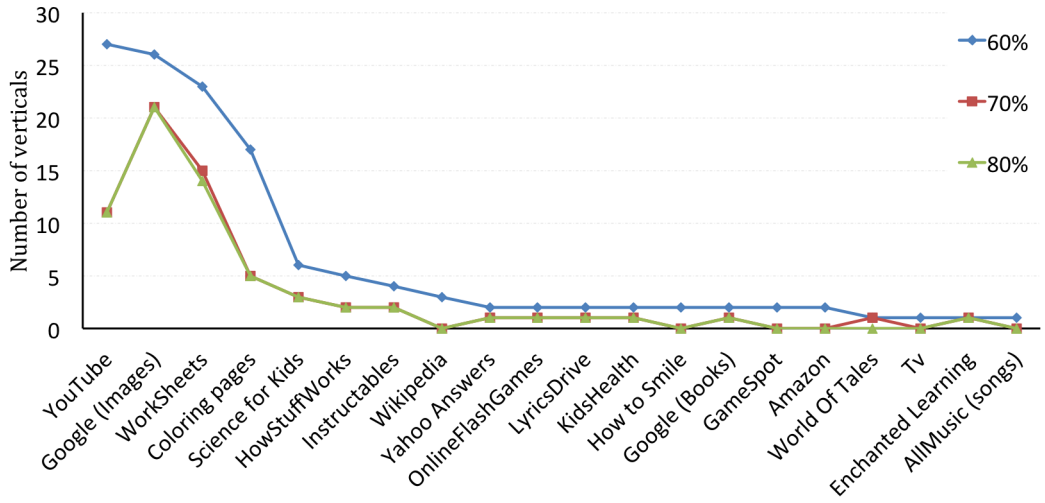


Figure 6.3: Frequency distribution of verticals for the first experimental protocol.

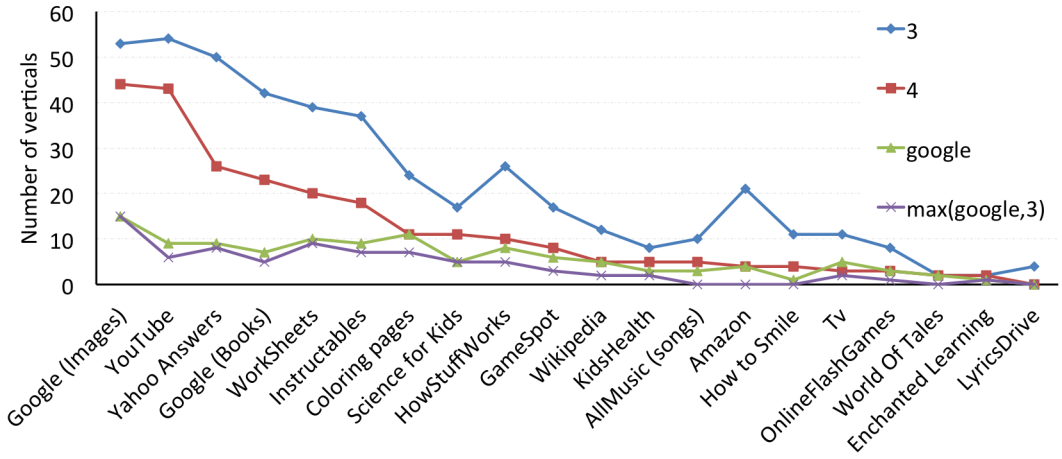


Figure 6.4: Frequency distribution of verticals for the second experimental protocol.

visual-oriented verticals are preferred, that is *YouTube*, *Google Images*, *Coloring pages* and *Worksheets* are the most frequent verticals assessed as relevant. This result may be due to the bias generated by visual content in the paired assessments. We believe that the exposition of visual content is more appealing to the user when they are asked to make an assessment against text based results. This hypothesis is also supported by the fact that the best performing verticals at higher threshold values are the image oriented verticals (*Google Images* and *Worksheets*). The bias towards visual oriented verticals has also been reported before in the literature [Arguello et al., 2012; Sushmita et al., 2009]. Anecdotally we also observed that children oriented educational websites (*e.g.* science for kids, instructables) were more frequently chosen

as relevant than *Wikipedia*, which is a relatively trusted source. All the other verticals (*e.g.* music, lyrics, games) were less frequent and were located in the bottom of the long tail distribution.

For the second experiment protocol a vertical is said to be relevant if the averaged score assigned by all the assessors to a (*query*, *vertical*) pair is greater that a given threshold. Figure 6.4 shows the distribution obtained when using the values 3.0 and 4.0 as threshold (recall that we used a graded system from 1 to 5 to judge each result set).

Additionally, we also employed as thresholds the averaged scores obtained by the *Google Web* vertical (on a query basis) and the maximum score between the *Google Web* vertical averaged core and 3.0 respectively. The last two thresholds were employed in order to make a fair comparison between the relevant verticals obtained with the two experimental protocols. Recall that in the first experimental protocol we compare each set of vertical results against the *Google Web* vertical, for this reason we employed the Web vertical score as threshold for the experiments.

In Figure 6.4 we observed similar distributions for the thresholds 3.0 and 4.0, although all the frequencies in the latter were lower as a consequence of the higher threshold value. Some of the verticals in the long tail also rank differently (e.g. *Amazon*, *KidsHealth*, *Tv*). Nonetheless, the most frequent verticals were the same when using both thresholds: *Google Web*, *Google Images*, *YouTube* and *Yahoo! Answers*. For the thresholds *google-score* and *max(google-score,3.0)* we observed large differences in the distribution in respect to the first two thresholds employed. Even though the top 2 most frequent relevant verticals were the same (*Google Web* and *Google Images*. Verticals such as *Youtube*, *Yahoo! Answers* and *Google Books* were not prominent as it was the case before. In general terms all the other verticals were assigned as relevant with less frequency. This result indicates that the score obtained by the *web vertical* was often higher that the score assigned to the other verticals given that the frequency of the relevant verticals are significantly trimmed when using *Google Web*'s score as threshold. It is important to mention that even though this is the case, in about 10% of the queries, *Google Web* was not chosen as relevant (when using as reference 3.0 and 4.0 as thresholds).

An important observation of the previous results is that in the second experimental protocol we did not observed the visual content bias observed with the first experimental protocol, as it was shown with all the thresholds. These results suggest that the second methodology can also simulate the first when using the *web vertical* score as threshold while avoiding the bias generated when comparing text results against visual content.

Inter-agreement metric	Score
average pairwise percent agreement	82.86%
fleiss' kappa	0.683
FK agreement	0.828
average pairwise cohen's kappa	0.682
krippendorff's alpha	0.683

Table 6.3: Inter-agreement scores found for the task.

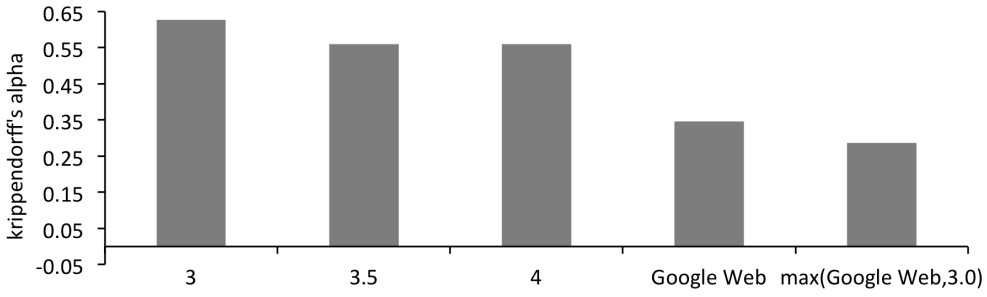


Figure 6.5: Inter-assessor agreement for the second experiment protocol for several thresholds.

6.5.2 Inter-assessor agreement

We analyzed the inter-assessor agreement for both experimental protocols to quantify the quality of the assessments under different relevant thresholds and to identify the relevant threshold values that maximize assessor agreement. The motivation is to use these thresholds in our experiments of vertical selection. We are also interested in investigating the threshold in which both protocols lead to a similar set of relevant verticals. For the first experimental protocol, the survey consisted of a sample of 90 queries, which lead to 3360 decisions (comparisons between the web vertical and each one of the other verticals). On average, three assessors evaluated in each pair. Table 6.3 shows the inter-assessor scores obtained for the experiment. We list the most common metrics employed in IR and natural language processing. All the metrics show a substantial agreement between assessors. This result indicates that assessors consistently interpreted the task and that they agreed in distinguishing content that is suitable for children in the age range specified.

For the second experiment protocol we measured the inter-assessor agreement by establishing that each pair *query, vertical* is assessed by n assessors (*e.g.* 3 coders) and the assessment is binary (relevant or non-relevant) based on the threshold defined in the previous section: 3, 4, web vertical score and *max(web vertical, 3.0)*. Note that this encoding is slightly different to the one employed in the first experimental

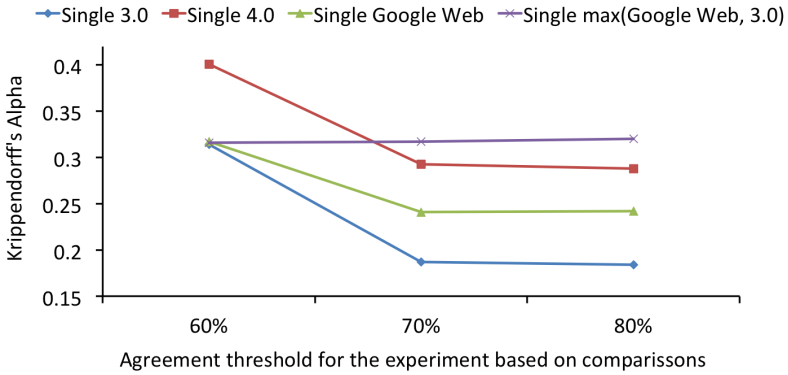


Figure 6.6: Agreement between the two experiment protocols.

protocol, in which we had three possible values: (relevant, non relevant (*web vertical* is preferred) and none of the two sets are relevant. Figure 6.5 shows the inter-assessor agreement using the Krippendorff's alpha score. We only report these results since the averaged Fleiss' kappa agreement obtained was almost identical to the Krippendorff's alpha scores. We found that at lower threshold values the inter-assessor agreement in higher, which is expected since lower threshold values represent a larger score interval to discern between relevant and non-relevant. It was observed that the lowest agreement is obtained when using the *web vertical* score as threshold. In general terms, the inter-assessor agreement was slightly lower that in the first experimental protocol.

Additionally, we compared the agreement between the lists of relevant vertical using both experimental protocols. The agreement was also measured using the Krippendorff's alpha score. In this case, we have two coders (the results of each experimental protocol) and the coding is binary: relevant or not relevant. We compute the score for all the possible threshold combinations of the two methods. Recall that in the first protocol the threshold refers to the percentage of users assessing a vertical as relevant while in the second protocol the threshold refers to the graded score (between 1 and 5). In this case, the cut-off is established by considering as non-relevant, the pairs in which the averaged scores of all the assessors are under the threshold defined. It is important to mention that we estimated the agreement scores under two scenarios: with and without considering the *web vertical*. We made the distinction because this vertical is not assessed directly with the first protocol since it assumes that the results from verticals are relevant only if they add value to the results provided by the *web vertical*. Nonetheless, we artificially created an assessment for this vertical by setting it as relevant if for a given query all the assessors prefer the *web vertical* at least in one of the paired comparisons.

Figure 6.6 and Figure 6.7 show the results obtained using these two modalities (*i.e.* with and without the *web vertical*) respectively. For the former the maximum

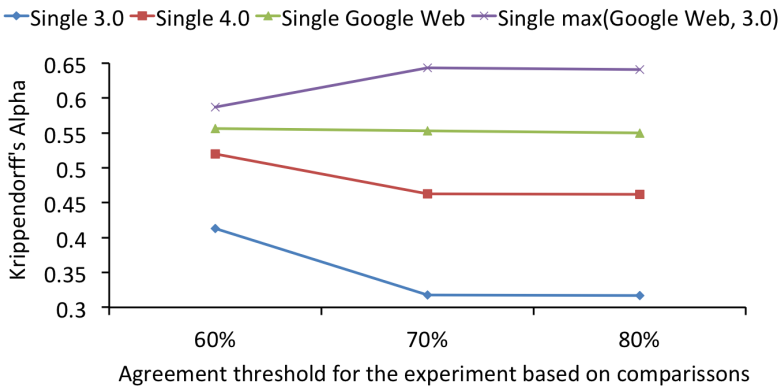


Figure 6.7: Agreement between the two experiment protocols (Including Google Web).

agreement score obtained was 0.41 using as threshold 4.0 for the second protocol and 60% for the first protocol. Scores between 0.4 and 0.6 are considered moderate agreement [Bermingham and Smeaton, 2009; Schaer, 2012]. Nonetheless we observed that the score values were more stable when we set the second protocol threshold as *max(web vertical, 3.0)*. This result is interesting because it suggests that by using the *web vertical* score we can simulate, at some extend, the vertical set obtained by the first protocol, having the advantage of avoiding the visual bias identified for the first protocol. In Figure 6.7, we observed larger agreement scores (maximum of 0.632). Similarly, higher values of agreement are obtained consistently when using the score of the Google Web vertical.

On overall, the previous results suggest that both approaches lead to relatively high inter-assessor agreement. However, the first protocol provides more consistent results since the inter-assessor agreement is higher. It is important to mention that the second protocol is not prone to visual bias and provide a wider set of relevant verticals per query, which is useful in the construction of exploratory information systems. In addition, we found for the second protocol that the threshold *60%* lead to the highest assessor-agreement, thus we will employed this threshold for our experiments (Section 6.8).

6.6 Vertical Size Estimation

The corpus size estimation is highly important to understand verticals' characteristics and quality. The size estimation of a corpus is also a key feature in the selection of search engines in federated search and distributed search [Thomas and Hawking, 2007; Xu et al., 2007]. In our scenario, the estimation of the vertical size is crucial since this statistic is needed in the best performing resource selection methods such

as *ReDDe*. Recall that the system for children we envisage has only access to the verticals through limited query interfaces in which is only possible to submit a query to receive a limited number of results. Under this non-cooperative environment, the search engines do not provide collection summaries from which global statistics about the vocabularies of the collection can be inferred, for this reason they need to be estimated from vertical samples. We will show how these estimations are used in vertical selection in Section 6.7.

Si and Callan [2003] proposed the use of the *capture-recapture* method through query-based sampling to estimate the size of the collections for the problem of resource selection in non-cooperative environments. The capture-recapture method has been used traditionally in the Ecology field for the estimation of the population size of species. It works as follows: a predefined number of animals are captured, marked and then released. After a certain amount of time, a second sample of animals is captured in the same area and the new sample is inspected to estimate the intersection between the two samples. The population is estimated using the sizes of the two samples and their intersection using the following expression:

$$s' = \frac{|sample_A| * |sample_B|}{|sample_A \cap sample_B|} \tag{6.1}$$

In search engines this process is carried out using query-based sampling: A set of queries is sent to the target search engine and the documents returned by the search engines are collected. This process is repeated and the estimation is based on the size of the number of documents collected in the two samples.

Shokouhi et al. [2006] explored the problem of resource selection in non-cooperative environments and proposed a query-based sampling *QBS* method based on the capture-recapture method [Si and Callan, 2003] referred as *multiple capture-recapture*. They showed with a large set of heterogeneous collections that their method outperforms previous approaches using search engines in non-cooperative environments. We employed their method to estimate the sizes of the verticals chosen for our collection. In their method, the *capture-recapture* process is repeated T times using samples of size m and the estimation is carried out by counting the size of the intersection of each pair of samples. Concretely the estimation is performed with the following expression:

$$s' = \frac{T(T-1)k^2}{2D} \tag{6.2}$$

where T is the size chosen for each sample, k is the number of samples and D is the accumulated number of duplicates found in the intersection within each pair. In our approach, we are not only interested in estimating the size of the collection but also in the size of the content available for the target domain in each vertical. We

employed the *multiple capture-recapture method* and a random samples of queries from the query set of our collection to estimate the size of the content of interest for children in each vertical. Recall that these queries were chosen to be representative of the topics of interest for children, as it was described in Section 6.3.1.

We carried out an analogous process to estimate the size of the verticals of content that is not oriented to children. For this purpose we employed a set queries known to be submitted to extract information in other domains (*i.e.* non for children). For this set we employed the same methodology described in Section 6.3.1, but instead of using the *Kids and Teens* seed urls we employed the global categories of Dmoz, which are not present in the categories for children. In this fashion, we obtained size estimations for each vertical of content that is oriented for children and non-children. The results of the estimates are shown in Table 6.4. These values were obtained by using a set of 2K queries. We set the parameters T and k to 50 and 25 respectively. The sample was constructed by choosing randomly 5 queries from the set of 2K queries (with replacement) and collecting the top 10 results for each query. Similar parameter values have been used in previous studies [Shokouhi et al., 2006]. It is important to mention that the set of 90 queries employed to gather user assessments were not employed in the samples generated for the size estimation process to avoid bias in the evaluation of vertical selection methods.

Consistently, we observed large ratios between the estimations using the *grown up* and *kids* queries with verticals known to be large and targeting all kind of public. For instance, the ratios for the verticals *web*, *question/answers*, and *images* were 34.0, 1.6 and 13.0 respectively. Inverse ratio trends were observed when considering verticals focused on children topics, such as the gaming and educational verticals.

The ratio found for the verticals *games online* and *education (livescience)* were 0.6 and 0.4 respectively. These results are interesting for resource selection because the ratios give us an estimation of the likelihood to find content for children in the verticals. In the following section, we will explore the use of these ratios along with the vertical size estimation in the problem of vertical selection.

6.7 Resource selection methods in IR for children

We employed two well-known vertical selection methods: *ReDDe* and *Clarity*. The former has been shown as one of the most effective methods for resource selection in federated search in cooperative and non-cooperative environments [Si and Callan, 2003]. More recently, this method was adapted for state-of-the-art aggregated search systems [Arguello et al., 2009] and it was proven to be one the most discriminative sources of evidence in vertical selection. It is important to mention that models built on query logs have been shown to provide higher performance than *Redde* [Arguello et al., 2009, 2010].

Vertical	Kids	Grown ups
web	803,297	27,377,757
games online	51,235	31,282
games	79,910	36,965
images	4,749,306	63,878,640
coloring pages	776,072	1,258,471
worksheets	931,287	1,006,777
books	963,956	30,115,533
question/answers	8,613,190	14,553,214
stories	3,928	3,478
shopping	562,308	22,485,899
music	121,920	540,501
videos	728,842	177,227
movies	61,443	131,131
encyclopedia	61,186	409,160
how-to	172,481	114,445
school(livescience)	11,666	5,308
school(howstuffworks)	3,267	3,674
school(discovery)	18,241	23,117
s. activities (sciencekids)	4,851	3,597
s. activities (howtosmile)	3,267	11,027
lyrics	455,305	161,271
health	7,185	2,939
tv	13,390	310,538

Table 6.4: Vertical size estimations using the set of *kids* and *grown up* queries.

However query logs are inaccessible in the settings of the aggregated search system we are interested in, since the verticals belong to third parties, contrary to the case of the aggregated search setting considered by Arguello et al. [2009]. For this reason, we employed in this work *ReDDe*, which is defined in equation 6.3.

$$ReDDe_q(V_i) = |V_i| \sum_{d \in R} I(d \in S_i^*) p(q|M_d) p(d|S_i^*) \tag{6.3}$$

where $p(d|S_i^*) = \frac{1}{|S_i^*|}$, $|V_i|$ is the size of the vertical, S_i^* is the set of sampled vertical documents and $p(q|M_d)$ is the query likelihood score of the document d. The score is estimated from an index combining the document samples of all the verticals. In this work we index the documents using the Terrier search engine[1] and score them using the Hiemstra's Language Model [Hiemstra, 2002].

The second baseline employed is *Clarity*, which was originally used as a measure

[1] http://terrier.org

of retrieval effectiveness. *Clarity* is defined in the following way:

$$Clarity_q(C) = \sum_{w \in V}^{|V|} p(w|M_q) log \left(\frac{p(w|M_q)}{p(w|M_c)} \right) \tag{6.4}$$

where $P(w|M_q)$ is defined as:

$$p(w|M_q) = \frac{1}{z} \sum_{d \in R} p(w|M_d) p(q|M_d) \tag{6.5}$$

where $p(w|M_c)$ is the probability of w, which is estimated using the collection language model and $p(w|M_c)$ is the query likelihood score of the document and it is estimated using a index of the documents of the target collection only: $p(q|M_d)$. The variable z is defined as $\sum_{d \in R} p(q|M_d)$. The documents are also scored using language models with linear interpolation smoothing [Hiemstra, 2002]. For *ReDDe* and *Clarity* the top 100 documents where employed to calculate the document scores.

We also redefine *ReDDe* to exploit the size estimations obtained with the children and non-children queries. The intuition is to use the ratio between these two size estimations to boost those verticals that contain more content available for children, since these verticals have a higher likelihood of providing content that is suitable for them. Equation 6.6 shows the new definition:

$$ReDDe'_q(V_i) = \frac{|V_i^{kids}|}{|V_i^{adults}|} \sum_{d \in R} I(d \in S_i^*) p(q|M_d) \tag{6.6}$$

where $|V_i^{kids}|$ and $|V_i^{adults}|$ are the size estimations obtained using the two set of queries. We will show in the next section that this definition lead to better performance with our test collection.

Representing verticals through social media

Nowadays, social media is widely used to describe and share web resources on the Internet. We believe that the descriptions provided by these thousands of users can be beneficial for the problem of vertical selection, particularly on specialized domain environments.

In this work, we utilize bookmarks from the social website *Delicious*[1], in which users can share bookmarks of their favorites websites by providing a list of describing tags. For instance, tags describing the domain *www.howtosmile.org* (such as *science*, *math*, *lessons*) can be used to emphasize this vertical if we are able to infer that the tags are related to the intent of the query (*e.g.* which is the case for the query *school*

[1] `http://delicious.com`

science fair).

For this purpose we create a tag representation of the query and the vertical. A language model is employed to assign a retrieval score to the verticals. For both representations we used the Delicious crawl collection built by Wetzker et al. [2008], which contains around 130 million bookmarks.

The query is represented as a bag of tags using the top *10* results of an index containing the documents in the Dmoz *kids and teens* section. The intuition is that the tags describing the top results from this index are a fair representation of the intent that the query has in the domain of information for children. Similarly, a vertical is represented as a bag of tags associated to the urls of the sample of documents extracted from the target vertical. It is important to mention that we associate each url to a set of tags by *(i)* finding the url in the collection provided by Wetzker et al. [2008], and by *(ii)* extracting tags from the title and snippet of the results by using the vocabulary of tags. The latter strategy was used given the low coverage of the collection for the small verticals.

Based on this vertical representation we rank the verticals for a query using a language model: $p(V_i|q)$, which is defined in the following fashion:

$$p(V_i|q) = \frac{p(q|V_i)p(V_i)}{p(q)} \propto p(t) \prod_{j=1}^{|q|} p(q_j|V_i) \tag{6.7}$$

$$p(q_j|V_i) = \frac{cf(q_j, V_i) + \mu\, p(q_j)}{|V_i| + \mu} \tag{6.8}$$

where $p(q_j)$ is the prior probability of q_j and μ is the Dirichtlet smoothing parameter. These probabilities are estimated using MLE on the artificially documents of tags created for each vertical and query.

We combine this probability score with the *ReDDe* score defined in Equation 6.3. For this purpose, we normalize *ReDDe* scores across verticals for each query and we us the following weighting scheme:

$$LMReDDe_q(V_i) = p(V_j|q) * ReDDe_q(V_i)^* \tag{6.9}$$

where $ReDDe_q(V_i)^* = ReDDe_q(V_i)/\sum_{k=1}^{|V|} ReDDe_q(V_k)$. An analogous definition was applied to *ReDDe-R* expressed in Equation 6.6.

6.8 Experimental Results and Discussion

We compared the performance of *ReDDe-R* (Equation 6.6), the social media language model (Equation 6.7), *LMReDDe* and *LMReDDe-R* (Equation 6.9) against the state-

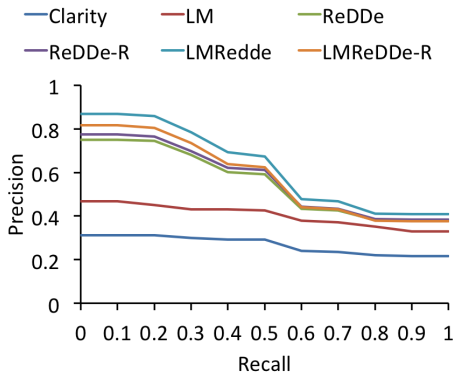


Figure 6.8: Protocol A (with *web vertical*)

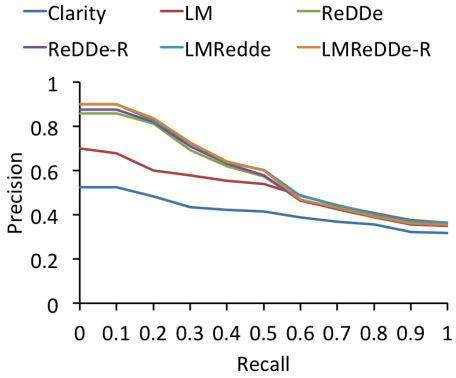


Figure 6.9: Protocol B (with *web vertical*)

of-the-art methods *ReDDe* and *Clarity*. For our experiments, we employed the 90 queries annotated with human assessments using both protocols: paired comparisons and single vertical assessments, which will be referred in our results as protocol A and B respectively. We employed as threshold the values 60% and 3.0 to set a vertical as relevant on the experimental protocols A and B respectively. For the language model we set experimentally the parameter μ to 2500. It is important to mention that other threshold values lead to consistent results

Figures 6.8 and 6.9 shows the precision and recall curves obtained for all the methods using the gold set obtained through the experimental protocol A. We found that *Clarity* is consistently outperformed by all the other methods and that the best performing methods are the ones based on *ReDDe*. We also observed that our three *ReDDe* variations outperformed both baselines being *LMReDDe* the best performing method for protocol A and *LMReDDe-R* for protocol B.

We noted that the *web* verticals is generally the first ranked vertical by most of the methods and that several queries in our test collection were only associated to this *web vertical*. Concretely, we found that from the 90 queries of the test set, 27 queries were associated only to the *web vertical*. For this reason, we decided to repeat our experiments ignoring this vertical from the collection. We believe that by ignoring this vertical we will have a clearer picture of the performance of the methods, especially on smaller verticals.

Figure 6.10 and 6.11 shows the precision-recall curves obtained by ignoring the effect of the *web vertical*. As it was the case before, *clarity* was outperformed by all the other methods. However we observed that *LM* outperformed all the other methods using protocol A, similarly with *LMReDDe-R* for protocol B. We believe this result is interesting because it shows the potential of using social media for vertical representation. Recall that this metric makes only used of social media tags

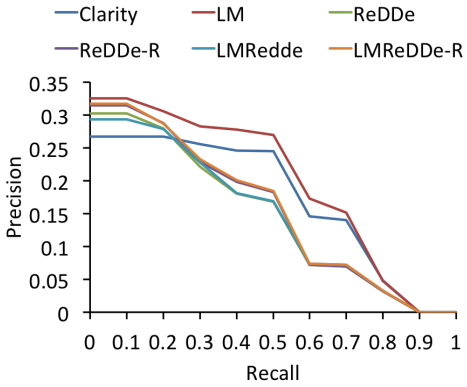


Figure 6.10: Protocol A (without *web vertical*)

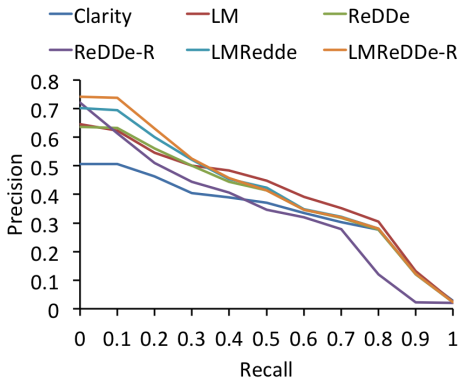


Figure 6.11: Protocol B (without *web vertical*)

Protocol	Web	Clarity	LM	ReDDe	ReDDe-R	LMReDDe	LMReDDe-R
A	Yes	0.242	0.375	0.527	0.548	**0.606**	0.564
	No	0.150	**0.179**	0.137	0.146	0.137	0.149
B	Yes	0.305	**0.588**	0.552	0.556	0.573	0.564
	No	0.287	0.377	0.360	0.382	0.382	**0.393**

Table 6.5: MAP results

to rank the verticals while all the other methods make uses of the entire content of the sampled documents.

Table 6.5 shows the MAP values obtained by all the methods with the two test sets and with and without the *web vertical*. The results are in line with the performance observed in the precision-recall curves. We verify the statistical significance of our results by comparing each pair of methods using the paired t-test for the equality of means with unequal variance. A statistical significance was acknowledged if the probability of the two means being equal (*e.g.* null hypothesis) is smaller than 5%. We found that all the differences were statistical significant except for the pairs: *ReDDe - LMReDDe* (using protocol A without the web vertical), *ReDDe-R - LMReDDe* (with *B* without the web vertical) and *ReDDe-R - LMReDDe-R* (with *A* and without the web vertical).

As a final remark, we observed that the models can behave differently according the test set. For example, the *LM* method seems to provide more gain for the test set created with the first protocol. We believe that further research is required to provide more robust mechanism to combine the scores of the different methods and to understand the scenarios in which each method is more beneficial.

topic	description
abyssinian guinea pig	Find any kind of information you can use to make a school presentation about these animals.
adding fractions	Find any type of content you can use to learn and practice how to add fractions (math)
butterflies	Collect content to build a presentation about butterflies for your classmates
crocodiles	Collect content to build a presentation about crocodiles for your classmates
easter	Collect content to explain a friend what easter is and things your friend can do with their family in easter

Table 6.6: Few of the topics employed in the game and their description.

6.9 Aggregated interface evaluation

So far we have described the methodology adopted to build a test collection for the evaluation of vertical selection methods in the domain of information for children, and we have proposed two methods to address this problem, considering the particularities of this information domain. In this and the following sections we addressed research questions *R.Q-6.3* and *R.Q-6.4* through a case study with a group of children and adults. We asked both group of users to engage a game designed to contrast the interaction in three types of pages: aggregated results from the verticals in our collection, Google aggregated results and Google simple results without image content.

In the game, users are asked to collect snippets from one of the three types of pages, given an open information task. In each game round a task is given in the form of a topic (which correspond to one of the queries employed in our test collection), and a description, which is given as a hint to help users understand the task. Table 6.6 presents few examples of the topics used. Users are allowed to select up to 5 results (i.e. maximum 5 clicks) from the list or skip the task. For both case studies a total of 35 topics were available for each page type.

Figures 6.12, 6.13 and 6.14 present the menu, task description scene, and main interface of the game. In the main menu users can start a new game, check the list of top scorers or watch a video tutorial explaining the idea of the game and its options[1].

[1]https://vimeo.com/72475434

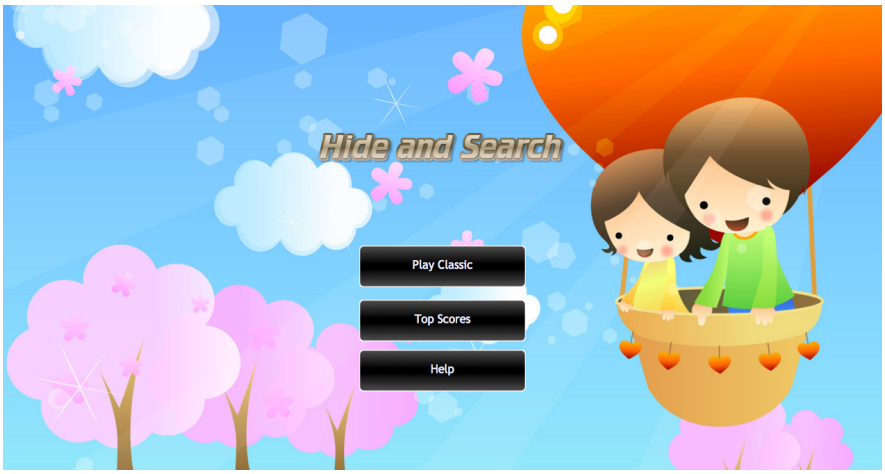


Figure 6.12: Menu presented to the users after they are logged in.

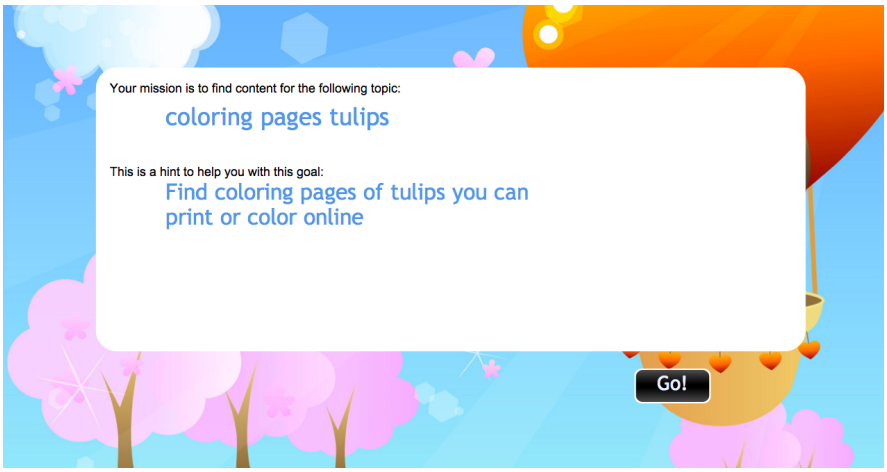


Figure 6.13: Task description scene presented to users at the beginning of each game session (i.e. game round).

Every time the user starts a new game round (i.e. task), a scene displaying the task and its description is shown. This scene was introduced to force users to read the information that they are asked to search. An example of how this scene looks like is shown in Figure 6.13. The game starts once the user finishes reading the task description and clicks on the next button. The main interface where the game takes place is shown in Figure 6.14.

On the top part of the screen is shown the control panel. In the control panel is always displayed *(1)* the task and a hint explaining the task in more detail. Left to the task is shown a star along with the number of clicks left that are available for the user *(2)*. A maximum of 5 clicks were allowed for each task. The number of clicks available is also shown below the description as a row of stars *(3)*.

Figure 6.14: Main interface where users are asked to select results given a topic. *(1)* task and description; *(2)* number of clicks available; *(3)* clicks available; *(4)* points gained in the session; *(5)* user name and accumulated points; *(6)* buttons to change the goal, access the main menu,and logout; *(7)* result page; *(8)* scroll buttons.

Next to the task description is found the number of points that the user has obtained in the session *(4)*. In this case with session we refer to all the activity registered while the user is logged in. In the right top corner is displayed the information about the user, that is the nickname (or *hero name* as it is called in the game) and the total number of points that the user has obtained throughout all the sessions *(5)*. Below this information 3 buttons are shown: *change goal*, *menu* and *exit game (6)*. The first button let the user to move to a different goal, independently of the number of clicks that are left in the current game round. This option is provided to avoid the situations in which users click on content that they are not sure to be relevant or that they may even dislike. If this situation occurs they can simply move to a different goal with content that they believe is more appropriate. The second button let the users go back to the main menu and the last button is equivalent to *logging out*.

Below the *control panel* is shown the result page in a frame, in which users click on snippets of information that they consider relevant of useful for the task provided *(7)*. The possible snippets that can be selected are limited and are shown while they explore the page with the mouse pointer or by touching the screen. In Figure 6.14 this functionality is exemplified with the green rectangle on the images. The block can be selected by clicking (or tapping) on the selected element.

Even though users can scroll with the standard mechanisms provided by modern web browsers, we added on the left side of the page prominent arrows that also provide the scrolling functionality *(8)*. We added this special buttons to motivate users to explore the content found in the lower sections of the page results, especially for the

case of children since they have been observed to ignore or simply not acknowledge this type of functionality [Druin et al., 2009].

6.9.1 Logging system

The main goal of the game is to learn the results and the verticals that are preferred by users of different ages. This goal is achieved by capturing the interaction of the users engaging the system.

Users are asked to create an account for the very first time they engage the game. We only ask for a nickname (i.e. hero name), password, age and gender. During the game we log all the *actions* carried out by the user. An *action* can refer to any of the following:

- New game round: This action occurs when the user is assigned a new task to solve. This action is triggered when the user starts a new game from the main menu (Figure 6.12), the user completes the maximum number of clicks allowed per game round, or the user clicks on the *Change* button to shuffle through the game tasks.
- Click: Clicks are allowed on the snippets or results available in each result page. The information associated to a click includes the task, page type, rank position, vertical, points gained, time-stamp and user.
- Game task shuffle: This action is triggered when the user clicks on the *Change* button. It assigns a new task to the user.
- Logout: This action is triggered when the users clicks on the *Exit* button. It ends the current session and sends the user to the login initial page.

Note that the privacy and identity of the user is not registered and cannot be inferred directly or indirectly from the actions logged.

6.9.2 Point system

The user is awarded a specific number of points after each click. The number of points awarded is estimated based on the number of users that have clicked in the target element. The results on which users agree more provided more points than results that are rarely clicked. This strategy is analogous to the one employed in *games with a purpose* (GWAPs)[von Ahn and Dabbish, 2008]. These games have increasingly being used in the literature to harvest human work to solve or complete specific computational problems. Games with a purpose differ from standard crowd-sourcing platforms in that the process of solving computational tasks is entertainment, and

in the sense that users do not engage the tasks with the motivation of solving a problem but with the intention of being entertained, which implies that users are not (necessarily) expecting a remunerated reward.

Two prominent GWAPs games developed by von Ahn and Dabbish [2008] are *ESP* and *Peekaboom*. In the former, two users are selected randomly and asked to label an image. Players are rewarded when they assign the same labels to the image. In this way users are encouraged to label as accurately as possible the image given. The game is entertainment because users compete to sum up points and because the time constraints added to the game. *Peekaboom* is a game designed to label images. In this game two players are selected randomly. One user is assigned the role of *Peek*, and the other the role of *Boom*. The *Peek* user is given a blank screen and has to guess what the image is about, using words (either verbs or nouns) as labels. The *Boom* user reveals parts of the image to guide the other user in the labelling process. The game is engaging because the cooperation nature of the game and the competitiveness involve in the process of gathering points.

Recently, similar efforts have emerged in the information retrieval field to automatize the process of collecting relevant judgments. For instance Eickhoff et al. [2012] studied the benefits of modeling this process by mapping the judgment of relevant documents to a labeling problem in which topics are assigned to a keyword. They concluded that the quality of the judgments collected are comparable to the judgments of the best performing TREC participants under the settings of the TREC 2011 Crowd-sourcing task. They showed that the cost of crowd-sourcing platform tasks can be reduced when using this type of games.

The key observations of GWAPs include the cooperative nature of the game to enforce quality in the feedback collected, the challenging nature of the game, which is introduced by the competitiveness between users and time constraints; and by rewarding the users for their effort using a point system. For our case, users are awarded points based on their agreement (to encourage quality), and the competition between users is motivated by providing a list of top scorers.

For each page we provided a maximum of M points, which are distributed among the N results available in the page. The score assigned to each element is proportional to the likelihood of that element being chosen by the target group of users. This probability is estimated based on the number of users that have clicked on the element. In other words, the probability of the elements can be seen as the probability distribution of each element being clicked. Formally, this probability is defined in the following manner:

$$p_{task}(result) = \alpha \frac{cf_u(result)}{cf_u(task)} + (1 - \alpha)\frac{1}{N} \tag{6.10}$$

and the score is defined using this probably:

$$score_{task}(result) = \lceil M \times p_{task}(result) \rceil \tag{6.11}$$

where $cf_u(result)$ and $cf_u(task)$ are the number of users that have clicked on the result and the total number of users that have engaged the task. Recall that a task refers to a specific page of results (any of the three types) for a given topic. The parameter α is employed to control how much importance has the agreement between players in the estimation of the score. If $\alpha = 0$ then every result is attributed the same number of points. If $\alpha = 1$ then only the elements with at least one click are attributed points. Note that for $\alpha = 0$, If there has been only one click in the task, then the element clicked receives the maximum number of points (i.e. M). In section 6.10 we describe the parameters adopted for the case studies.

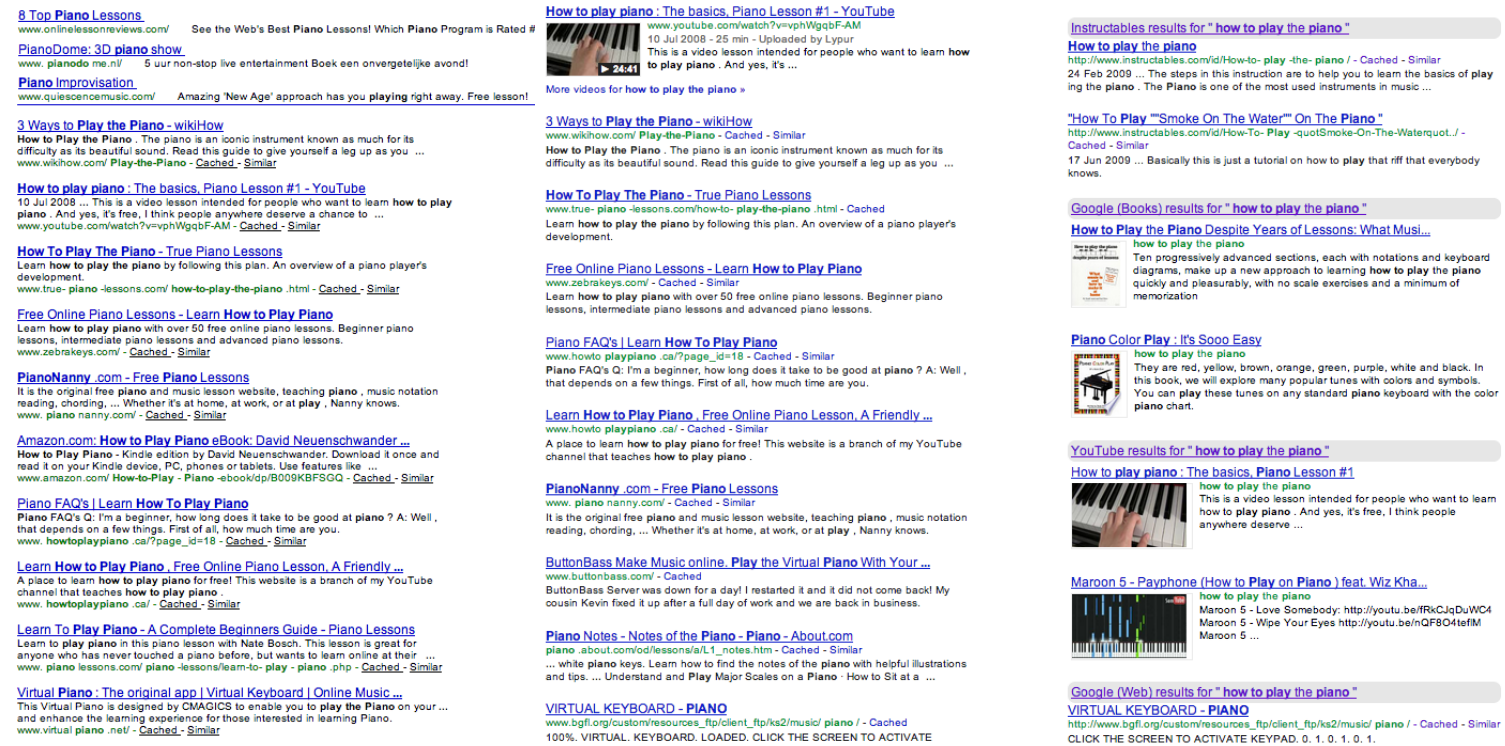
Additionally, to increase the difficulty of the game, we identified manually, in the three types of pages, results that were not relevant for the query. Only, clearly irrelevant content was flagged as irrelevant. This was also the case for the advertisements displayed in the pages that were all flagged as irrelevant. Users were taken away three points if they clicked on one of these results. On average each result page had two irrelevant results (including the advertisements).

6.9.3 Page types description

We will refer to the three types of pages as *plain*, *google* and *aggregated*. Figure 6.15 shows the three types of pages available in the game. The *plain* page is built using the results of the Google Custom Search[1]. This service provides results from the web vertical of the Google search engine. No images or news are provided in these results. Nonetheless, results from multimedia verticals, such as *YouTube* can still be shown in the result list. The same is the case for advertisements. An example is shown in Figure 6.15(a). The *google* page is built using the state-of-the-art Google Search[2]. The difference between this and the *plain* page type is the inclusion of images, news and thumbnails in the results. Figure 6.15(b) exemplifies this type of page.

The *aggregated* page (Figure 6.15(c)) is built using the results fetched from the verticals selected in Section 6.3.2. The page was built artificially using at most 4 verticals besides the *web* vertical. The verticals were chosen from the relevant judgments describe in Section 6.5. Particularly, we employed the assessments using the single page grading approach, in which each vertical was evaluated in isolation using a graded system from 1 to 5.

[1] `http://www.google.com/custom`
[2] `http://www.google.com/`



(a) Plain (b) Google (c) Aggregated

Figure 6.15: Example of each page type for the topic *how to play the piano*. The *Google* and *Aggregated* examples were truncated due to space constraints.

We employed the top 4 verticals scored by the assessors. This is, on average, the maximum number of verticals that are employed by state of the art search engines and that has been used in previous studies [Arguello et al., 2011].

Only verticals with a score greater than 3 were allowed (averaged across the score given by all the assessors). For each vertical we included the top two results returned by the vertical. For the case of image oriented verticals (i.e. coloring pages, worksheets and images) we arranged the top 5 results in a single row, which is the case in state of the art search engines. Thus, the results from these verticals count as 1 in the list of 10 results. For the case of the web vertical we used as many results as necessary to complete the 10 results per page. This number varies across topics because not all the topics were assigned more than 3 vertical as relevant. Only topics in which the *web* vertical was chosen as one of the relevant verticals were included in the case study.

For the case of the *aggregated* page type, we included for each topic the advertisements (if any) found in the equivalent *Google* page. We also designed the page using the style presented by the *Google* page in order to avoid biases towards the stylistic aspects of the pages. The only noticeable change is the header introduced for the vertical results. The header presents the name of the vertical (e.g. *How To Smile results for the query...*). Nonetheless the changes are subtle as can be observed in figures 6.15 and 6.15(a).

6.10 Case study settings

Two case studies were carried out: the first with a group of children and the second with a group of adults. The first case study was carried out in an international elementary school in the Netherlands. Even though participants were from different ethnic groups and nationalities, they were all highly proficient with the English language. For the second group we employed the *CrowdFlower* platform.

6.10.1 Elementary school group

Students enrolled in the sixth grade of an international primary school in the Netherlands were addressed. In the Netherlands, the primary school is composed of 8 years and children in the sixth grade are typically aged 9 to 10. The case study consisted of a 2 hour session with the children in the classroom. Prior to this session, parental consent forms were sent to the parents of the children. The form summarized the purpose of the game and the privacy policies adopted for the collection of data. Parents were encouraged to be part of the process by letting their children engage the game online from their homes. The class addressed was composed of 16 students (approximately), 11 of them were allowed to be part of the study.

At the beginning of the session, children were introduced to the game by showing the steps required to create an account and log into the system. The dynamic of the game was explained along with a small tutorial about how to play and how to use the controls provided in the interface. Emphasis was given in the goal of the game (finding results relevant to the goal presented to them). After children did not have more questions about the game, we moved to the next step. It is important to mention that to add an extra factor of motivation and competitiveness, a prize was offered to the 3 players who scored the maximum number of points.

In the next step, children were called in small groups. They were allowed to play the game for 15 minutes after they were assisted to create an account. The process was repeated for all the other children able to join the experiment. At the end of the session children were shown the list of top scorers and the prizes were awarded.

Table 6.7 summarizes the age, gender and task completion characteristics of the users (first column for the children group). 65.6% of the users were male. Nine years old was the median age of the group. On average 2.7 users completed each tasks and each user completed on average 26 tasks.

6.10.2 CrowdFlower group

We created a task in CrowdFlower using the *external hit* approach. For each work unit, workers were provided instructions about the task and a link that pointed to an external server where the unit was carried out, in our case to the location of the game. A token was provided once the worker finished with the work unit successfully.

Few modifications were introduced to the interface in order to avoid spammers and detect sloppy workers. A unit of work was defined as the completion of 4 game rounds (i.e. 4 game tasks). As it was mentioned before, negative points were awarded when the user clicks on advertisement or on one of the results known to be incorrect or irrelevant for the task, a maximum of 1 click in any of these elements were allowed throughout the 4 game rounds. Thus, workers were not given a token at the end of the session if they made more than one mistake. Additionally, we slightly modified the functionality of the *Change button*. We allowed workers to shuffle through the tasks with this button, however tasks in which this function was utilized were not counted as completed. We opted for this decision to avoid workers skipping tasks without clicking or exploring the result page. At the same time we did not remove the button altogether to avoid the situation in which users may feel forced to click on results even if they are not certain about the relevancy of these results. For each work unit was paid 3 dollar cents. Only workers from the United States were allowed to work on the task.

On overall, 63 users engaged the task and their average age was 30 years old. A slightly larger number of women engage the crowd-sourcing task (52.3% vs 47.6%)As

	Children	CrowdFlower
num. users	11.0	63.0
avg. age	9.3	30.5
median age	9.0	29.0
male	63.6%	47.6%
female	36.3%	52.3%
avg. users per task	2.7	8.3
avg. tasks per user	25.9	15.0
avg. users per query	8.1	24.8
avg. queries per user	25.9	15.0

Table 6.7: Users characteristics and average task completion for the group of children and *CrowdFlower* users. The last three columns show the percentage of *CrowdFlower* users that report being the guardian of a child in the target age range.

it was expected, a larger number of users per task (in respect to the school of children) was observed as a consequence of the large number of users and the greater time that they had available. Nonetheless, it is important to mention that a maximum of 15 tasks were allowed for each user. In this way, we reduced the chance of users engaging the same tasks. The characteristics of this group are shown in Table 6.7.

6.10.3 Parameters tuning

Two parameters are required to be adjusted for the game: the smoothing parameter α from the point system in Equation 6.10, and the number of clicks allowed in each task. The parameters were tuned experimentally by running a pilot study using CrowdFlower. Initially, we disabled the *Change* button to guarantee that users employed all the clicks that were available. We set the number of allowed clicks to 3 and the α parameter to 0.5. During the case study we realized that users were not fully exploring the result list and most of the clicks were registered in the first three ranked positions. We also observed that the agreement between users was not noticeable with the α value chosen.

Based on this observation we increased the number of allowable clicks to 5 and the λ parameter to 0.7. We observed that these settings lead to more exploration and a more rewarding experience for the users. Similarly we enabled the *Change* button to avoid false positive clicks.

6.11 Log analysis results

Three cleaning steps were applied to ensure that the results were not prone to bias caused by learning effects. Learning effects may arise if the users engage the same query more than once during the game (e.g. same query with different page types).

		Children	CrowdFlower
Clicks	plain	0.122	0.428
	google	0.127	**0.436**
	aggregated	**0.204**	0.348
Verticals	plain	**0.673**	**0.783**
	google	0.515	0.585
	aggregated	0.372	0.478

Table 6.8: Krippendorff's alpha agreement for each group of users. The table summarizes the agreement found in terms of the content clicked and the verticals selected.

Similarly, the first task engaged by the user was considered as a trial task, in which users get familiar with the dynamics of the game. This was particularly the case for children. The cleaning steps are summarized as follows:

- The first task engaged by each user was ignored.
- Queries that had been done previously by the user in a different page type were ignored.
- Tasks that were done previously by the user (i.e. same query and page type) were also ignored

Most of the results reported are micro-averages. We also computed macro-averages in respect to the user, as it was the case in Chapter 3 and 4. We only reported the formers given that the trends observed with both approaches were very similar. We will clarify for each analysis if this is not the case.

Comparisons between averages across page types and between the two groups of users (children and *CrowdFlower* workers) were carried out. The two-tailed t-test is applied and a difference is considered statistical significant if the probability of the null hypothesis is smaller than 0.5%. P-values will be reported along each analysis. Additionally, the Krippendorff's alpha coefficient was utilized as a statistical measure of inter-rater agreement. This statistic allows an arbitrary number of coders and incomplete data. The same coefficient was utilized in Section 6.5.2.

6.11.1 Assessor Agreement

The agreement in terms of the content clicked (and the verticals selected) was measured using Krippendorff's alpha coefficient. The encoding for the analysis can be seen as a matrix with m rows and n columns, where m is the total number of results (on average 10 per query) presented to the users, and m is the maximum number of coders that addressed each result in the page. Thus, each row in this matrix correspond to one of the results of a particular task. Missing values were allowed in

Page type	Agreement
Plain	0.244
Google	0.314
Aggregated	0.358

Table 6.9: Vertical selection agreement between user groups for each page type, measured with the Krippendorff's alpha agreement score.

the matrix. Only units that were judged by at least 3 assessors were included in the analysis. One query was discarded for the analysis using this restriction. Table 6.8 summarizes the results.

Even though the click agreement found for the group of children was not high for the three page types, the largest agreement was found for the case of the *aggregated* page type (0.204 against 0.122 and 0.127 for the *plain* and *google* page type). In general, the agreement was larger for the case of the *CrowdFlower* users. In this case the *google* pages had the largest agreement (0.436 against 0.348 and 0.428 for the *aggregated* and *plain* page types).

Larger agreement values were observed when considering vertical selection instead of individual page clicks. For both groups, the *plain* and *google* pages had the largest agreement. This result is partly due to the smaller number of verticals available in this type of interfaces. For instance, most of the results available in the *plain* pages are classified into the *Web* vertical. On the other hand, each *aggregated* page has results from 4 verticals, typically. Moreover, there are up to 15 types of verticals across topics. We believe the greater number of high quality verticals led to a larger variability in the selection of verticals for this type of pages. We also measured the vertical selection agreement between the two groups of users. For this purpose we mapped the clicks to vertical relevant judgments. A vertical shown in a given task is considered relevant, if more than half of the users that engaged the task clicked on any result from that vertical. We only included units with at least 3 assessors for the group of children and 5 assessors for the case of the *CrowdFlower* users. Note that in this case the coding consists of a matrix with two columns (two coders), and the rows are the available verticals for each task. Table 6.9 presents the results. We found that the two groups agreed more on the selection of verticals for the case of the *aggregated* page.

Nonetheless, we observed that for all the pages the agreement, although fair for the case of the *aggregated* type, was not substantial. In order to identify the verticals in which the disagreement occurred, we quantified the proportion of agreement for each vertical, and we broke down this percentage to identify the number of cases in which the group of children consider a vertical relevant (through their clicks) while the second group did not, and vice-versa. Tables 6.10, 6.11 and 6.12 present the results found for the *aggregated*, *plain* and *google* page types, respectively.

		Agreement Breakdown			Disagreement breakdown		
Vertical	Agreement	K (Yes) - C (Yes)	K(No) - C (No)	Disagreement	K (Yes) - C (No)	K(No) - C (Yes)	Queries
Rotten Tomatoes	0.0%	0.0%	0.0%	100.0%	0.0%	100.0%	1
Google (Books)	16.7%	100.0%	0.0%	83.3%	40.0%	60.0%	6
Instructables	45.5%	0.0%	100.0%	54.5%	50.0%	50.0%	11
GameSpot	50.0%	0.0%	100.0%	50.0%	0.0%	100.0%	2
Wikipedia	58.3%	14.3%	85.7%	41.7%	0.0%	100.0%	12
Yahoo Answers	60.0%	0.0%	100.0%	40.0%	75.0%	25.0%	10
Google (Web)	70.6%	70.8%	29.2%	29.4%	0.0%	100.0%	34
HowStuffWorks	71.4%	60.0%	40.0%	28.6%	50.0%	50.0%	7
Google (Images)	76.2%	0.0%	100.0%	23.8%	80.0%	20.0%	21
YouTube	80.0%	0.0%	100.0%	20.0%	33.3%	66.7%	15
WorkSheets	88.9%	0.0%	100.0%	11.1%	100.0%	0.0%	9
Coloring pages	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%	4
Amazon	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%	2
Science for Kids	100.0%	100.0%	0.0%	0.0%	0.0%	0.0%	5
KidsHealth	100.0%	66.7%	33.3%	0.0%	0.0%	0.0%	3
AllMusic (songs)	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%	2
How to Smile	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%	2
OnlineFlashGames	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%	1

Table 6.10: Vertical selection agreement details between children and *CrowdFlower* users for the Aggregated page type. The columns K(Y) and C(N) refers to the proportion of queries in which the children group (K) considered a vertical relevant while the*CrowdFlower* users did not (i.e. C(N)).

	Agreement Breakdown			Disagreement breakdown			
Vertical	Agreement	K (Yes) - C (Yes)	K(No) - C (No)	Disagreement	K (Yes) - C (No)	K(No) - C (Yes)	Queries
Instructables	33.3%	0.0%	100.0%	66.7%	25.0%	75.0%	6
Wikipedia	45.5%	80.0%	20.0%	54.5%	0.0%	100.0%	11
News	66.7%	0.0%	100.0%	33.3%	0.0%	100.0%	3
YouTube	66.7%	20.0%	80.0%	33.3%	20.0%	80.0%	15
Google (Web)	72.7%	100.0%	0.0%	27.3%	22.2%	77.8%	33
Science for Kids	80.0%	0.0%	100.0%	20.0%	0.0%	100.0%	5
Amazon	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%	1
Rotten Tomatoes	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%	1
Yahoo Answers	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%	1
KidsHealth	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%	2

Table 6.11: Vertical selection agreement details between children and *CrowdFlower* users for the Plain page type.

	Agreement Breakdown			Disagreement breakdown			
Vertical	Agreement	K (Yes) - C (Yes)	K(No) - C (No)	Disagreement	K (Yes) - C (No)	K(No) - C (Yes)	Queries
Instructables	40.0%	0.0%	100.0%	60.0%	0.0%	100.0%	5
YouTube	42.1%	50.0%	50.0%	57.9%	63.6%	36.4%	19
Google (Images)	42.9%	11.1%	88.9%	57.1%	91.7%	8.3%	21
Rotten Tomatoes	50.0%	0.0%	100.0%	50.0%	100.0%	0.0%	2
Science for Kids	50.0%	50.0%	50.0%	50.0%	100.0%	0.0%	4
KidsHealth	50.0%	100.0%	0.0%	50.0%	0.0%	100.0%	2
News	60.0%	0.0%	100.0%	40.0%	50.0%	50.0%	5
Yahoo Answers	66.7%	50.0%	50.0%	33.3%	100.0%	0.0%	3
Wikipedia	66.7%	50.0%	50.0%	33.3%	0.0%	100.0%	12
Google (Web)	87.9%	100.0%	0.0%	12.1%	0.0%	100.0%	33
Amazon	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%	2
HowStuffWorks	100.0%	100.0%	0.0%	0.0%	0.0%	0.0%	1
AllMusic (songs)	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%	1

Table 6.12: Vertical selection agreement details between children and *CrowdFlower* users for the Google page type.

The values are sorted in ascending order in respect to the agreement percentage (first column). The agreement and disagreement percentages are breakdown to report the details of the discrepancies between the two groups. For instance the column *K (Yes) - C (No)*, refers to the percentage of queries in which the majority of user from the children group (denoted as K) selected results from a given vertical while the group of *CrowdFlower* users (denoted as C) did not. To ease the interpretation of results we added in the last columns of the tables the number of queries in which the vertical was presented.

For the verticals *Google Books*, *Instructables*, *Yahoo Answers* and *Wikipedia* it was found that the two groups disagree in more than 40% of the queries. No clear trend was observed for the verticals *Google Books* and *Instructables*. For instance, *Instructables* results were clicked by children for around 50% of the queries for the cases in which the disagreement occurred, while adults did not. However, for the case of the *Yahoo! Answers* we found that children clicked on results from this vertical for 75.0% of the queries in which there was disagreement, and *CrowdFlower* users did not click on this type of results. For the *Wikipedia* vertical it was observed the opposite behavior, the *CrowdFlower* users clicked on this type of content for all the queries in which there was disagreement while children did not.

Higher agreement was observed for the verticals *Google (Web)*, *HowStuffWorks*, *Google (Images)*, *YouTube* and *WorkSheet*. For these verticals the two groups agreed on more than 70% of the queries. For *Google (Web)* and *YouTube* we found that the *CrowdFlower* users tended to click on results from these verticals while children did not click on these results when there was disagreement between the two groups. The opposite trend was observed for *Google Images* and *Worksheets*.

On overall, we observed high agreement and large number of clicks for both groups of users on educational services with moderated content for children, such as *Science for Kids KidsHealth* and *HowStuffWorks*. Even though, relatively large agreement values were observed for the *Google (Web)* results, children seen less keen to click on these results (similarly for the case of *Wikipedia*). The opposite behavior was observed for the case of *Google (Images)* since children tended to click more on these results than adults. Interestingly this was also the case for the vertical *Yahoo Answers*, in which children clicked more than the adult group of users.

Tables 6.11 and 6.12 show that results from the verticals *KidsHealth* and *Science for Kids* were less noticed by both group of users. On the other hand, consistent trends were observed for the *Google (Web)* vertical. A different behavior was observed for the verticals *Google (Images)*, *Youtube*. In these verticals both group of users had low agreement and children tended to click more on the results from these two verticals. These results suggested that it was easier for children to identify the content that is especially designed for them (e.g. *KidsHealth* and *Science for Kids*) in the *aggregated* interface.

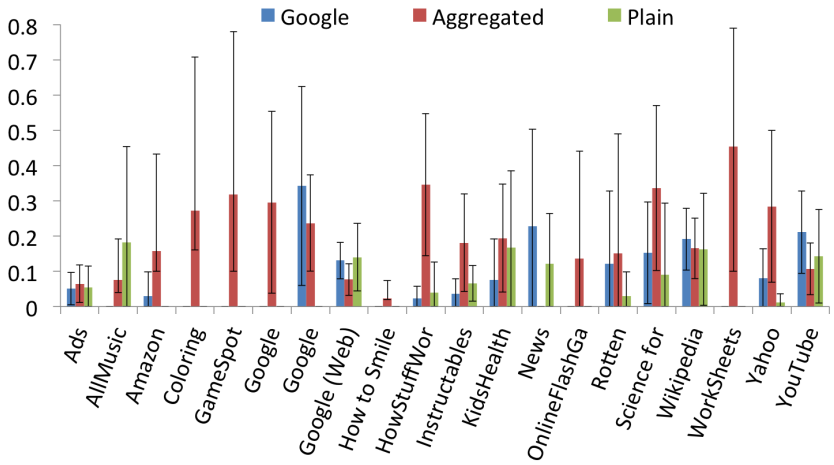


Figure 6.16: Likelihood of clicking on a vertical in each page type for the set of child users.

6.11.2 Vertical selection distribution

We estimated the likelihood of clicking on a vertical result for each page type. The motivation is to compare the vertical preferences of both age groups and the likelihoods across the page types. The likelihood of clicking on a vertical is estimated in the following fashion:

$$p(v) = \sum_{i=0}^{|U|} p(v|u_i)p(u_i) \tag{6.12}$$

$$p(v|u_i) = \frac{cf(v,u_i)}{cf(u_i)} \tag{6.13}$$

where $cf(v,u_i)$ is the number of queries in which the user clicked on at least one of the results of vertical v. The value $cf(u_i)$ is the total number of verticals shown to the user. This value is calculated by summing up the number of verticals of all the tasks engaged by the user. The probability $p(u_i)$ is assumed to be uniformly distributed. Finally, U refers to the set of users that engaged the game.

The results for the children and CrowdFlower users are shown in Figure 6.16 and 6.17, respectively.

Child users results

Figure 6.16 shows that the likelihood of selecting a vertical varies depending of the page type engaged. For the *aggregated* page type, the most likely verticals to be selected were: *WorkSheets*, *HowStuffWorks*, *Science for Kids*,*GameSpot*, *Google*

(Books) and *Yahoo Answers* Note that the top 4 verticals provide results that are moderated for children. For the *google* page type the most popular verticals were *Google (Images)*, *News*, *YouTube*, *Wikipedia* and *Science for kids*, nonetheless the likelihood values found for these verticals were lower than the values observed for the *aggregated* page type. For instance the likelihood for the *Science for Kids* vertical in the *aggregated* pages was 0.30, and for the *google* pages was 0.15. For the case of the *plain* page type, we found that *KidsHealth*, *Wikipedia*, *Youtube* and *Google (Web)* were the most likely verticals to be clicked.

The high likelihood for the *News* vertical came as a surprised since we did not observe an important volume of clicks on this type of content in Chapter 3, which was the finding that led us to exclude this vertical from our set of verticals. Nonetheless, exploring the results provided by this vertical, we found that up to 50% of the clicks were on results that were not relevant for the task.

Interestingly, verticals such as *Google (Images)* and *YouTube* were less likely to be clicked on the *aggregated* pages. This is due to the high quality content for children provided by the other verticals such as *HowStuffWorks* and *WorkSheets*.

We found statistical significant differences (as described in Section 6.11) for the following combinations of vertical and page types. P-values are reported in parentheses: *aggregated-google* (0.005) and *aggregated-plain* (0.007) for the vertical *HowStuffWorks*; *aggregated-google* and *aggregated-plain* (0.028) for *Google (Books)*; *aggregated-plain* (0.049) for *Science for Kids*; *aggregated-google* (0.045) for *Instructables*; *aggregated-plain* (0.018) and *aggregated-google* (0.041) for *Yahoo Answers*; *aggregated-plain* (0.004) and *aggregated-google* (0.0006) for *Google (Web)*; *aggregated-google* (0.041) for *YouTube*.

CrowdFlower users results

As it was the case with the group of children, the *CrowdFlower* users had different vertical preferences according the page type. The results are summarized in Figure 6.17. For the case of the *aggregated* pages, the verticals *Google (Books)*, *Science for Kids* and *Wikipedia* had the highest likelihood of being clicked. For the *plain* page type the preferred verticals were *Wikipedia*, *Google (Web)*, *YouTube*. The verticals *Google (Images)*, *Wikipedia*, *Google (Web)* were the most likely verticals to be clicked for the *google* pages. On overall we observed that verticals with moderated results are preferred in the *aggregated page* and that popular verticals such as *Google (Web)*, *Google (Images)* and *Youtube* were less likely to be clicked on the *aggregated* pages. Nonetheless, we still observed large likelihoods for verticals such as *Wikipedia* and *Instructables* across all the page types. This behavior was not observed for the group of children.

Most of the results were statistical significant given the large amount of *Crowd-*

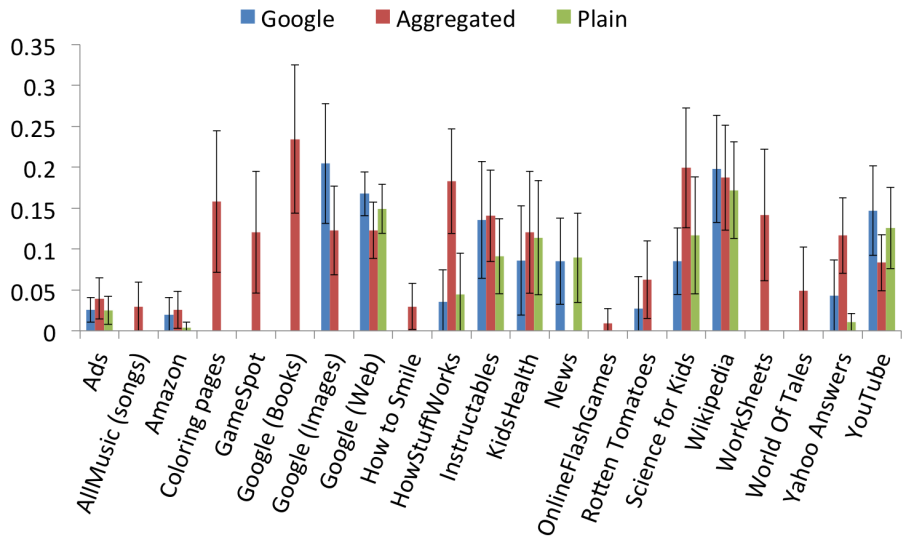


Figure 6.17: Likelihood of clicking on a vertical in each page type for the set of CrowdFlower users.

Flower users that engaged the game: *aggregated-plain* (0.01) for *Rotten Tomatoes)*; *aggregated-plain* (0.002), *aggregated-google* (3E-5) for *Google (Web)*; *aggregated-plain* (9E-4), *aggregated-google* for *HowStuffWorks*; *aggregated-plain* (3E-05), *aggregated-google* (0.021) for *Yahoo Answers*; *aggregated-plain* (0.001), *aggregated-google* (0.001) for *GameSpot*; *aggregated-plain* (0.048), *aggregated-google* (0.048) for *AllMusic*; *aggregated-google* (0.007) for *Science for Kids*; *aggregated-plain* (0.037), *aggregated-google* (0.037) for *How to Smile*; *aggregated-plain* (3E-06), *aggregated-google* (2E-06).

Comparison across age groups

The likelihood of clicking on each vertical is contrasted for each page type between the two groups of users in Figures 6.18, 6.19 and 6.20 for the *plain*, *google* and *aggregated* page types, respectively.

Figure 6.18 shows that the CrowdFlower users had a slightly higher likelihood to click on several verticals, such as *Google (Web)*, *Wikipedia* and *Instructables*. This group also had a higher likelihood to click on v verticals providing moderated content for children such as *Science for kids*. On the other verticals the trend is reversed. For most of the verticals the differences were marginal. Similar trends were observed for the *google* pages (Figure 6.19). Nonetheless, for the latter, children were more likely (than adults) to click on the *News* vertical and image oriented verticals such as *Google (Images)* and *YouTube*.

In general for the *aggregated* pages (Figure 6.20) verticals with moderated content were more noticeable to children and the likelihood of clicking on results from these

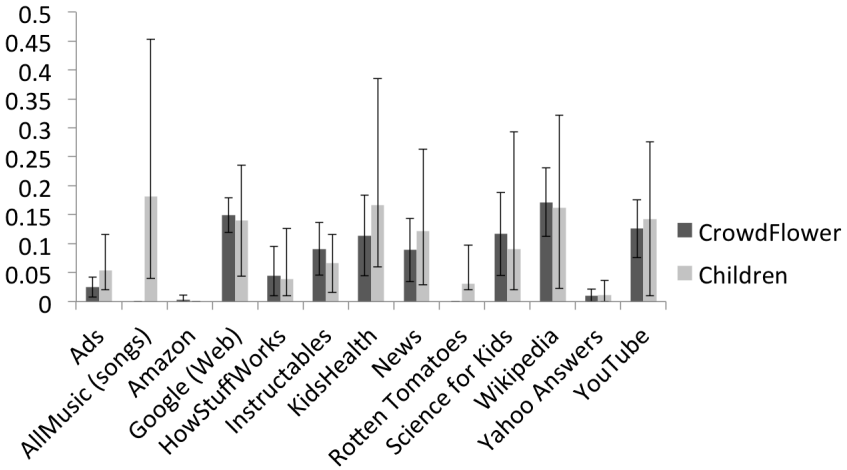


Figure 6.18: Vertical likelihood for both groups using the *plain* page.

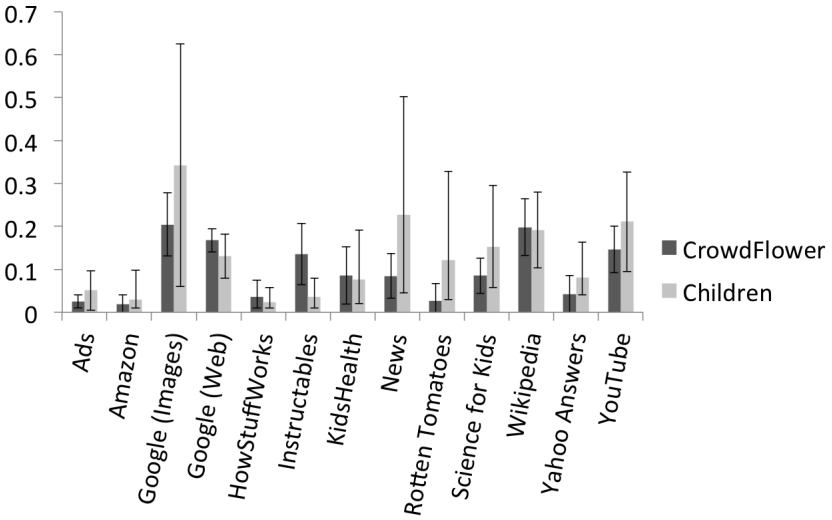


Figure 6.19: Vertical likelihood for both groups using the *google* page.

verticals were larger than the click likelihood of adults (i.e. *KidsHealth*, *HowStuff-Works*, *Science for Kids*). We also observed that the likelihoods of clicking on *Google (Images)* and *YouTube* results were lower on this page type, particularly for children. We consider this a positive results because it suggests that children were more prone to click on text based results over visually based results, as a consequence of a better selection of text results from the moderated verticals.

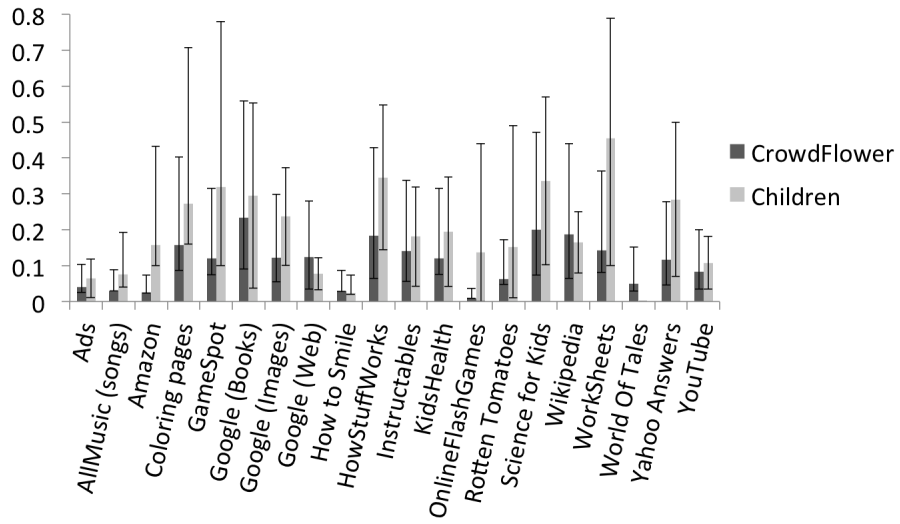


Figure 6.20: Vertical likelihood for both groups using the *aggregated* page.

	Children	CrowdFlower
Plain	4.4	3.9
Google	5.0	4.1
Aggregated	4.5	4.5
	p-values	
Google-Plain	0.831	0.001
Aggregated-Google	0.023	0.001
Aggregated-Plain	0.022	1.7E-07

Table 6.13: Average rank positions and p-values using paired t-test at 95% confidence level for each page type pairing.

6.11.3 Rank distributions

Similar rank averages were observed for all the pages. Table 6.13 summarizes the results. On average children (and *CrowdFlower* users) clicked on high ranked elements (top of the list) in the *plain* pages. The opposite trend was observed for the *google* and *aggregated* pages for both group of users. Thus, users explored less the list of results in the *plain* pages.

Even though the differences were small between rank averages, we found that in all cases the differences were statistical significant (using the paired t-test at the 5% level of confidence), except for the difference between the *plain* and *aggregated* pages for the group of children, p-values are reported in Table 6.13 as well.

It is important to mention that the results reported in Table 6.13 disregard the clicks on advertisements, which are located on the top of the result pages. A noticeable difference between the two groups is the proportion of clicks on this type of

	Children	CrowdFlower
Plain	1.54	1.30
Google	1.57	1.43
Aggregated	1.69	1.48
	p-values	
Google-Plain	0.600	1.1E-04
Aggregated-Google	0.021	0.175
Aggregated-Plain	0.007	3.5E-07

Table 6.14: Average number of points awarded per result clicked and p-values using the paired t-test at 95% confidence level for each page type pairing.

content. Around 3% of the clicks of CrowdFlower users were on advertisement, and 10% for the case of children.

Even though, users were penalized in different ways for clicking on ads (CrowdFlower users were allowed to make at most one mistake), this result indicates that children struggle to identify advertisements. This behavior was consistent across the three page types.

6.11.4 Interaction analysis

In the following paragraphs we explore the usage of the *Change button*, the average number of clicks, the number of points gained by the users, and the average time spent per tasks and per click for each group of users.

Points awarded

Table 6.14 presents the average number of points gained on average for each click. A large number of points suggests high agreement between the users. Recall from Section 6.9.2, that results that are more frequently clicked lead to a greater number of points. It is important to clarify that the results presented do not exclude the negative points awarded to users when they clicked on ads or results known to be irrelevant. Nonetheless, the results obtained when these clicks are excluded are consistent with the results reported.

For both group of users, a greater number of points was awarded, on average, in the *aggregated* interface. For the children group of users, this result is consistent with the agreement coefficient reported in Section 6.11.1. Even though, the differences were not large, all the results were statistical significant except for the paired comparison between the average scores estimated in the *google* and *plain* result pages.

We expect larger point differences between page types as more users engage the system, since the estimation of points is heavily dependent of the number of users providing clicks.

	Children		CrowdFlower	
	Result	Task	Result	Task
plain	8.9	33.5	5.9	21.3
google	7.7	31.6	5.6	18.4
aggregated	9.2	36.1	6.3	23.8
	p-values			
Google-Plain	0.044	0.625	0.360	0.052
Aggregated-Google	0.019	0.292	0.028	0.016
Aggregated-Plain	0.668	0.561	0.186	0.306

Table 6.15: Click and game duration. The values are in number of seconds.

Clicks per game round

Children tended to skip the *plain* and *google* pages more often than the *aggregated* pages. The following were the percentages of clicks on the *Change* button: 19.2%, 17.5% and 14.6% for the *plain*, *google* and *aggregated* pages, respectively.

On overage, there were only marginal differences in the number of clicks on each page type. For instance the average number of clicks (averaged per number of tasks engaged) in the *plain*, *google* and *aggregated* page types were 3.4, 3.5 and 3.3, respectively. None of the paired comparisons were statistical significant.

Although we observed that CrowdFlower users also tended to use the skip button less in the *aggregated* page types, the perceptual differences were not large and non-statistical significant: 1.9%, 5.1% and 0.4% for the *plain*, *google* and *aggregated* pages, respectively. A different behavior was observed in respect to the average number of clicks. In general, the average number of clicks per game round was larger (4.46, 4.56 and 4.48) than the average number of clicks accounted for children. However, we believe this difference is due to the restriction added to the game in CrowdFlower. Recall that for these users all the clicks available had to be employed in order to count the game round as completed.

Click and game duration

The game round duration accounts for the moment in which the user click on the button *Go!* in the task description scene to the last click registered in the game round (or the click on the *Change* button). Click durations are estimated based on the number of seconds between the click event on the result, and the previous event (either another click or the event representing the start of the game round). Game rounds that were skipped, or that did not register any click were ignored in the results reported. Similarly, clicks that took longer than *100* seconds were ignored. This threshold was employed in Section 3.4.4 (Chapter 3) to define *long clicks*. Nonetheless, the trends were consistent when using longer thresholds. The

	Plain	Google	Aggregated
HowStuffWorks	-	9.5	14.2
YouTube	10.1	6.2	10.7
KidsHealth	5.0	9.3	8.7
GameSpot	-	-	8.3
Science for Kids	10.0	5.5	8.1
Google (Web)	8.6	7.9	8.1
Instructables	6.0	11.5	7.8
Yahoo Answers	4.0	8.3	7.4
Wikipedia	7.4	5.3	7.2
Ads	4.7	7.2	6.8
Google (Images)	-	4.9	6.8
Google (Books)	-	-	5.3
WorkSheets	-	-	5.0
Rotten Tomatoes	10.0	6.5	4.0
OnlineFlashGames	-	5.3	4.0
Coloring pages	-	-	4.0
Amazon	-	10.0	3.5
News	5.5	8.8	-

Table 6.16: Click duration on the verticals in each page type for the child users. Verticals without sufficient data are excluded from the table (e.g. Coloring pages in the *plain* page type). Values sorted in descending order using the *aggregated* pages values.

click and game duration are summarized in Table 6.15.

On overall, the time spent on the *aggregated interface* was larger for both age groups. Significant statistical results were found for the *CrowdFlower* group of users. This result suggests that users were more engaged with the results presented on these pages.

The average click duration trends were similar for both age groups. We observed that clicks were longer on the *aggregated* pages. We believe this result also suggests a greater engagement with this type of results. An alternative interpretation is that children need a longer time to parse and understand the elements of the *aggregated* pages, however given that several of the results were from moderated verticals and designed for the language capabilities of children, we opted for the first interpretation. Moreover, clicks from Crowdflower users were also longer in the *aggregated* pages. In this case statistical results were found for both group of users when comparing the *google* and *aggregated* pages. Even though the trends were similar for both age groups, children spent more time per click and per game round.

Additionally, the average click duration per vertical was estimated to have a better understanding of the time spent to click on different type of results (e.g. image or text based). The estimation is analogous to the click dwell time explained before. Tables

	plain	google	aggregated
Rotten Tomatoes	-	13.5	7.0
World Of Tales	-	-	6.3
Yahoo Answers	2.5	3.4	6.2
KidsHealth	3.2	1.6	5.3
Ads	6.5	4.9	5.2
Wikipedia	4.0	4.0	5.1
Instructables	4.5	3.4	4.6
YouTube	4.9	4.6	4.5
Amazon	-	4.5	4.5
Coloring pages	-	-	4.4
HowStuffWorks	2.4	1.7	4.4
WorkSheets	-	-	4.4
Science for Kids	4.6	3.5	4.3
GameSpot	-	-	4.2
Google (Web)	4.5	3.8	4.1
Google (Books)	-	-	4.0
OnlineFlashGames	-	-	4.0
Google (Images)	-	3.9	3.7
AllMusic (songs)	-	-	3.2
News	3.4	3.3	-

Table 6.17: Click duration on the verticals in each page type for the CrowdFlower users. Verticals without sufficient data are excluded from the table (e.g. Coloring pages in the *plain* page type). Values sorted in descending order using the *aggregated* pages values.

6.16 and 6.17 summarize the results. On overall, we observed for the case of children (Figure 6.16) that clicks on most of the image based results, such as *Coloring Pages* and *Google (images)*, took from 3 to 7 seconds. This was also the case for most of the results with thumbnails (e.g. *Amazon*, *Google Books*). For the text based results the time varied from 7 to 9 seconds. The exceptions were the *HowStuffWorks* and *YouTube*. Clicks on these verticals were longer than 10 seconds.

For the case of the *CrowdFlower* users (Figure 6.17 we observed that clicks on image based results varied from 3.5 to 4.5 seconds, while clicks on text based results varied from 4 to 6.3 seconds. Results with thumbnails where in the same range observed for the image based results. No statistical significant differences were found across page types.

These results were expected since image results can be judged faster than text based results. Interestingly, clicks on results with thumbnails seem to have a longer duration than the clicks on images, and shorter than the clicks on text based results.

6.12 Survey study results

A survey was delivered to the children that participated in the case study. Children filled the survey with the help of their teacher. The survey is shown in Appendix B. Ten out of eleven children filled the survey. The motivation of the survey was *(i)* to gather the overall impression of the children towards the game, and *(ii)* find out if children noticed differences between the three types of pages.

The first two questions of the survey were: *Did you find the game hard?* And *What did you like and did not like about the game?* On overall all the children liked the design of the game and they all found it fun and entertaining. Seven children did not find the game difficult. The other children reported that it was *a bit* difficult because the time restriction imposed, and because points were taken away when irrelevant content was clicked. Two children found that there were not many topics available in the game and three of them complained about the login page of the game.

During the study we noticed that the login and registration process was difficulty for this group of children. Both processes were designed using standard web forms, like most web services available online. Special care should be given to the design of the front page of web services aimed at children in order to increase their accessibility. Other alternatives to the standard web forms to capture information are worth to explore, for instance the usage of avatars.

The next question was: *Did you notice differences in the pages displayed in each game session?* Eight children did not notice any difference in the pages. They only noticed that the topics were different in each game round. Two children noticed differences, although only in the sense that some pages had more visual content than others, thus these two users were also unable to clearly distinguish between the page types. Children were also asked explicitly if they noticed that certain results were grouped together (making reference to the vertical results presented in the *aggregated* pages). Most of them replied that only images and YouTube videos were grouped together. These answers are interesting because reflect that even if children did not recognize the changes between the pages displayed, their behavior and interaction differed throughout the pages types for several of the metrics addressed in the previous sections.

6.13 Discussion of the case study results

To answer research question *R.Q-6.3*, we collected evidence showing that children and adults explore more aggregated pages with results from the moderated verticals than result pages from a state-of-the-art search engine and plain result pages without thumbnails or images.

We found that, on average, children and adults clicks were longer in the *aggregated* pages. The results for both groups were statistical significant when comparing the *google* and the *aggregated* page types. Similarly, the average time spent on each game round was longer in the *aggregated pages*. Given that the these pages present several results from moderated services that are known to be suitable for children, longer game round duration suggests more engagement and exploration. This is also reflected in the proportion of times the *Change* button was used to shuffle tasks. Children used more this functionality in the *plain* and *google* pages, which indicates that users were less certain about the relevancy of the results or less satisfied with the results provided by the latter two types of pages. A similar behavior was found for the *CrowdFlower* users. On the other hand, we only found marginal differences in the average rank position between the three page types and the results were not statistical significant.

It is important to mention that children were also found to agree more on the content clicked on the *aggregated* pages, which shows that for this group of users was easier to identify relevant and non relevant content in these pages. However, CrowdFlower users agreed more in the *google* pages.

The likelihood of clicking on a result from each one of the verticals was estimated to address research question *R.Q-6.4*. We also estimated the overall agreement of the verticals selected by mapping the clicks of the children and CrowdFlower users to vertical judgments. The best agreement is obtained with the *aggregated* pages, which suggests that the vertical relevant judgments gathered using CrowdFlower were suitable to design aggregated pages for children (this result was also reflected in the click agreement values). As we pointed out in the previous sections, even though this agreement was fair, it was not substantial. We explored in detailed the verticals in which both groups of users disagreed. *Google (Books)*, *Instructables Yahoo Answers* and *Wikipedia* were the verticals with the least agreement (under 60%). We found that children tend to find useful the results from *Yahoo Answers* more often than adults, and the other way around for the case of the vertical *Wikipedia*. These findings can be used to improve the quality of the aggregated pages. For instance, the disagreements can be seem as a biased introduced by adult judges and methods to take this bias into account are worth to be explored to create aggregated pages closer to the actual judgments of children. We also observed that most of the verticals providing moderated content for children *e.g. KidsHealth, Science for Kids* were among the verticals with the best agreement.

In terms of the verticals preferred by children across the set of topics, we observed that the verticals most likely to be clicked were the ones with educational content for children (e.g. *Science for Kids*, *Worksheets*), and the verticals *Google (Books)* and *Yahoo! Answers*. On overall, we observed that verticals results were more likely to be clicked on the *aggregated* pages. Greater likelihood values were observed for this page

type in respect to the *google* and *plain* page types, and in respect to the likelihoods observed for the *CrowdFlower* group of users. Similarly, children were less likely to click on *Google (Web)* results in comparison to adults.

This is an important result because it provides evidence that children interact more with the results presented in the *aggregated* pages. Interestingly, we observed that image oriented results (e.g. *Google (Images)* were less likely to be clicked on this type of pages while educational content had a higher likelihood to be clicked.

6.14 Conclusions and future work

We presented a detailed description of a simple methodology to build a collection for the problem of vertical selection in the domain of content for children. We contrast two methodologies to gather relevant assessments using CrowdFlower. We proved that the two methods lead to a different set of relevant verticals and that the former is prone to visual bias. We show that the different sets obtained by these methods can also lead to differences in the performance of vertical selection methods. We believe that the choice of either methodology is highly dependent of the targeted aggregated search system. For instance if the web vertical is always displayed it may be more beneficial to employ the paired comparison method since it has higher inter-assessor agreement. Nonetheless, further refinements are needed given the visual bias obtained by this method. We found that tags from social media are an effective resource for the problem of vertical selection given that was the best performing feature for several of the experimental settings. Similarly, the ratio between the sizes estimations (*i.e.* children and grown ups) led to a significant performance gain.

In regard to the evaluation of the aggregated pages, we found that children explored more the pages showing results from our test collection. Clear evidence of more interaction was observed in terms of the time spent on each page, the proportion of times the tasks were skipped, and the likelihood of clicking on results from educational and moderated verticals. We also observed that children agreed more on the content selected in the *aggregated* pages. Nonetheless, we observed that the vertical selection agreement between the two groups of users was low. This result suggests that the level of interaction and engagement can be highly improved by accounting for the discrepancies between the relevant judgments provided by the two groups of users.

6.14.1 Future work

There are several directions for future work. We showed a simple language model to rank verticals using tags from social media. However, more sophisticated methods to exploit these resources are worth to be explored (*e.g.* random walk, semantic

latent models). Features based on readability measures and the sentiment expressed in the text results are yet to be explored. Features derived from the search behavior of children are also promising given the noticeable differences between children and adults. Behavior information derived from search log sessions has been employed to improve the ranking quality of web results [Agichtein et al., 2006].

In this work, it was not explored the correlation between verticals and the impact of the context in which the vertical results are presented. For instance, if the top three ranked verticals are *Worksheets*, *Science for kids* and *Coloring pages*, it may convenient for some tasks to drop *Coloring pages* given their similarity to the results provided by the *Worksheets* vertical, leaving room to other type of results (e.g. videos). Improving the diversity of results [Zhu et al., 2011], particularly on exploratory informational tasks, is also worth to be explored in future research.

Other potential follow up studies involve evaluating the impact of the order in which the verticals are shown to the user, the number of verticals displayed in each page and the coherence of the information aspects presented by the vertical results. Exploring methods to improve the agreement between the relevant judgments of children and adults represent another important line of research. For instance, allowing only parents to judge the relevancy of the verticals may lead to better results given their closer (and more recent) familiarity with children. The usefulness of aggregated pages and particular moderated services also need to be explored under close defined and complex information tasks.

Chapter 7

Conclusions

This thesis explored in depth the search behavior of young users from 7 to 18 years old. A large scale study was carried out using search and toolbar logs. The aim was to understand the interaction of young users with search engines and their activities on a web browser. The key findings from this study allowed us to propose two solutions to improve the search experience of young users on the Internet. Firstly, query recommendation methods were proposed. Query recommendations help this user segment to focus their search on query aspects (i.e. query senses) that are of interest to and appropriate for them. The second solution goes a step further. Verticals (i.e. search services) are suggested to help children improve their chances of finding appropriate content. The last part of this thesis provided evidence of the benefits in terms of interaction and exploration for users aged 8 to 10 years old carrying out exploratory search tasks using blended results from a carefully selected set of verticals, specifically for children aged 8 to 10 years old. The following sections answer the research questions drawn up in Chapter 1, based on the findings shown throughout the chapters of this thesis. The last section of this chapter summarizes possible research lines for future work.

7.1 Searching content for children

Search sessions from the AOL search logs were carefully selected and analyzed to determine the search differences between users, targeting content for young users and for the average web user: *R.Q-1.1. What are the differences in search behavior of users targeting content for young users in respect to the average web user?*

In Chapter 2, we found clear differences between the search sessions of the users targeting content for young users (both children and teenagers) and the average web user: Longer queries, sessions with longer duration, more usage of natural language in the queries, higher number of clicks and a greater amount of clicks on low ranked

results in the case of young users, were some of the key log metrics that provided evidence of the difference (statistically proven) between the group of young users and adults. We also observed that the behavior of users searching information for teenagers was closer to the behavior of those searching content for children, and not closer to the behavior of the average web user.

From the results presented in Chapter 2 is not possible to guarantee that the results observed actually correspond to the average web behavior of children and teenagers. However, these results have important implications: they exposed the difficulty of finding content appropriate for young users in state-of-the-art search engines. For instance, the use of longer queries (in respect to the average web user) shows the need of using focused queries to retrieve high quality content for young users. This difficulty is also expressed by the larger amount of clicks on lower ranked results. We validated that the results obtained are not biased by the fact that the users were identified through the Dmoz directory. We showed that the search behavior of users targeting other information domains (e.g. finance) greatly differ from the behavior of the users targeting content for young users and from the average web user.

Conclusion 1: *Users searching for information for children and teenagers behave differently in terms of query and search session characteristics, when they are compared to the average web user. Even though the group, children and teenagers, are also statistically different, they behave similarly based on the query log metrics adopted. The differences found show clear difficulties, during the search process, to find content appropriate for children and teenagers.*

Our second aim was to determine if the topics targeted with the AOL queries are representative of the topics targeted by young users: *R.Q-1.2. Can we identify a representative distribution of topics of interest to young users in the Web through a set of queries aiming at content for them?*

In Chapter 2, it was verified qualitatively that the cue words of the AOL queries are representative of the topics searched by children and teenagers. Recall that cue words refer to those keywords that represent the different information senses of a given query. For instance, the cue words, *movie*, *toys* and *images*, are possible cue words for the query *cars*. The verification was carried out by clustering the cue words in central topics and contrasting them against the topics of interest to young users identified in previous case studies by Ofcom [2010].

We also addressed this research question quantitatively by estimating the Pearson's correlation between the topic distribution of the queries from the AOL logs, and a large scale set of queries submitted by young users (using the Yahoo! Search logs). This analysis is described in Chapter 3.. The distributions were obtained by mapping the queries to categories in the Yahoo! Directory. We found high correlation values

when pairing the distribution of queries according to the targeted age. The *kids* set of queries on the AOL logs, identified with urls targeting content for users between 8 to 12 years old, correlated best with the topics of queries submitted by users aged 6-7, 8-9 and 10-12 years old to the Yahoo! search engine. For *teenagers* AOL queries, the highest correlation was found when using the topic distribution of the queries of users aged 13-15 and 16-18 years old from the Yahoo! Search engine:

Conclusion 2: *The topics of queries identified through the Dmoz urls in the AOL search logs correlate with the topics targeted by queries submitted by young users. Thus, the query set identified in the AOL logs are representative of the topics searched by children and teenagers in a large scale search engine.*

7.2 What and How children search on the web

We hypothesized in Chapter 3, that young users struggle to find information on the Web, since we found in Chapter 2 that users in the AOL logs had difficulties in finding information suitable for children. A large scale sample from the Yahoo! Search logs was employed to characterize the search behavior of users from 7 to 18 years old. Special emphasis was given to the search difficulties of this user segment: *R.Q-2.1. Do young users struggle to find information with a large scale search engine, and how is this struggle reflected in their search behavior from a query log perspective?*

Recall, that we only employed log activity from registered users with reported age, enabling us to characterize the actual search behavior of young users on a large scale, which was a limitation of the study shown in Chapter 2.

Clear evidence of search struggle was found in the queries and search sessions of young users. Shorter queries and greater usage of natural language (compared to adult users) showed that young users had difficulty formulating specific queries using keyword based queries. Recall, that in Chapter 2, it was found that short queries reduced the chances of finding appropriate content and that, for children, relatively long queries are needed to access high quality. However, in Chapter 3 we observed that the queries submitted by young users are shorter than the queries of adult users, which suggests that young users are less likely to find content that is designed for them.

This problem was also reflected in the larger amount of short clicks in the sessions of young users in comparison to adult users. Similarly, a larger number of abandoned sessions were observed for the former group of users. These results showed that young users abandon web results quickly. This result can be explained by a sense of frustration during the search process, given that young users abandoned their sessions more often than adults and that the average number of clicks was significantly shorter

than the average found for adult users.

A prominent behavior observed for the group of children and teenagers was the tendency to click on top ranked results, and in general on elements located in the top areas of the result page. This bias was expressed not only in less exploration of the result page but also on a greater amount of clicks on advertisements. We also found that children had a higher likelihood of clicking on content that is unsuitable for them by accident (e.g. adult content), which can also hamper their search experience.

Conclusion 3: *Signs of search struggle were clearly reflected in metrics such as query length, click duration, session abandon rate, number of events per session and rank distribution. These results suggest frustration during the search process (by abandoning search sessions) and disorientation (clicks on advertisements, clicks on adult content by accident). The limited exploration of the result page shows that young users are not getting the most out of the information that is available to them on the Internet.*

The differences between age ranges defined for children, teenagers and adults were also analyzed: *R.Q-2.2. Does the search behavior and search difficulties of children, teenagers and adults differ in a large scale search engine (Yahoo! Search)?*

Overall, clear differences were found between the group of young users (children and teenagers) and adults in most of the query and session metrics explored in Chapter 3. Nonetheless, the differences between the groups of children and teenagers were often small. Bigger differences were found between users aged 16 to 18 and users aged 19 to 25, than between users aged 10 to 12 and 16 to 18. Metrics in which this was not the case were; query vocabulary size, usage of natural language and topic distribution. For instance, users below 12 years old had a significantly smaller vocabulary size compared to teenagers. Teenagers also used natural language in queries significantly more often than children and adults. On the other hand, clear increasing trends were observed for query and session length metrics (it was lower for young users).

Conclusion 4: *Clear differences were found in the search behavior between the groups of children (7-8, 9-10 and 10-12), teenagers (13-15, 16-18) and adults (above 19), based on the query log metrics explored. The search behavior of teenagers was closer to the search behavior of children. The jump to adulthood in terms of search behavior occurred in users aged 19 to 25 years old. Nonetheless, marked differences between children and teenagers were observed in terms of topic interests, query vocabulary and usage of natural language.*

In Chapter 3, it was hypothesised that stages of development are reflected in search

queries: *R.Q-2.3. Can we retrace stages of children and teenagers development, in terms of the topics they are interested in, through their queries and the characteristics of these queries?*

Aspects of development such as; topic interests, sentiment expressed in the queries, vocabulary size and readability level of the content clicked, were explored to address this research question. A clear correlation was found between the topic distribution and the age of the user. The correlation differences between the group of young users and adults were more pronounced when using a set of *how-to* queries. Similarly, certain topics were found to have a clear correlation with the queries of users in certain age ranges. For instance, the topics that correlated the best with the youngest group of users were: *games*, *arts* and *pets*. As it was mentioned above, smaller vocabulary sizes were observed in the queries submitted by children below 12 years old, which reflects the limited vocabulary of these users in respect to teenagers and adults. This behavior was also observed in the amount of clicks on content that is easier to read (i.e. basic readability level according the Google search classification). Our analysis of the sentiment expressed in the queries was not in line with the other results and were not conclusive. We expected to observe a greater amount of sentiment expressed in the queries of young users. However, no clear trends were found. Further research is required to understand the impact of the sentiment expressed in the content clicked on by young users.

Conclusion 5: *Stages of development were clearly reflected in the search queries submitted by children and teenagers. Concretely, topics were correlated to specific age ranges and the biggest correlation gaps were found between the topic distributions of the youngest and oldest groups of users. Consistent behavior was observed for the analysis of query vocabulary and the readability of the content clicked.*

7.3 Browsing activities of young users

Chapter 4 explored the activities that users carried out on the web browser. We were particularly interested in the activities, on the browser, that occur before the submission of a query on the web and multimedia search services. A large sample from the Yahoo! Toolbar logs was utilized, using a similar methodology, to group the log activity of users in meaningful age ranges.

Firstly, the amount of browsing activity was quantified to give a better understanding of how children and teenagers use web browsers. Additionally, the amount of multimedia search (search of images and videos) was also measured through the toolbar logs. Remember, it is not possible to carry out this estimation with the standard search logs we employed in Chapter 3: *R.Q-3.1. What activities are carried out*

by young users on the web browser besides web searches?, How prominent is browsing for each age range? At what ages are multimedia searches preferred?

In this work the term *browsing activity* refers to any kind of internet activity on a web browser that is not carried out within a search engine. Overall, we found a significantly larger proportion of browsing activity by young users, particularly by teenagers. For instance users aged 13-15 and 16-18 were browsing almost 9 times more than searching, while for adults the ratio of browsing versus searching was 5:1. Children below the age of 12 (years old) browsed 7 times more than they searched. This result suggests that state-of-the-art search engines need to make a greater effort to attract young users.

In Chapter 4, it was observed that the teenager group of users used multimedia search the most in comparison to the other user groups(13-15 and 16-18). Teenagers were found to submit queries to these types of services (or this type of service) twice as often as adults. The proportion of queries submitted by children (users younger than 13 years old) to multimedia search services were slightly lower compared to the adult segment. This result shows the difficulty of the youngest user groups to identify and access these search engines, rather than suggesting their lack of interest in the use of multimedia services. This was shown in the case study in Chapter 6, in which children clicked on a large proportion of results from multimedia services when result pages with multimedia results were displayed to them.

Conclusion 6: *Teenagers and children carried out a significantly larger amount of browsing activity on the Internet in respect to the amount of searching. Even though adult users also performed more browsing than search activities, the margin for young users is significantly larger. Teenagers were the group to submit the largest proportion of queries to multimedia search services. The low usage of these services found for users younger than 13 years old provided evidence of the need for improving the presentation of results from these verticals for this segment of users.*

Chapter 4 explored the browsing activities that occur before a search and that are likely to trigger a search: *R.Q-3.2. Which types of search and browsing activities are more likely to trigger searching the web and multimedia search engines in the case of young users? Do these triggers differ from those observed in adults users?*

A large percentage of browsing activity led to web search events in all age groups, nonetheless the proportion was slightly higher in the adult segment of users (30% for users under 19 and 34% for users over 25). Also, a significant amount of web searching is carried out at the beginning of a browsing session, this behavior was more accentuated in the case of child users (18.7% for users under 12 years old and 14% for users over 25 years old).

Most *browsing events* had a relatively low probability of triggering web search and

the differences between the age groups were, although statistical significant, marginal. Nonetheless, the *knowledge browsing category* (i.e. clicks on Wikipedia) showed different behavior and a high likelihood of triggering web search events. Larger differences between age groups were also observed. The likelihood observed for this event was around 11% for most teenagers and adults. For the group of children the likelihood was around 16% This result reflects the importance of providing appropriate educational content to the group of young users, particularly children. We addressed this issue in Chapter 6, in which we showed that children had greater click agreement and high likelihood of clicking on, educational content that has carefully been designed for their language capabilities.

In regard to the search activities that trigger multimedia search, *web search* events had the greatest likelihood of leading to a multimedia search query. Particularly, for cases in which the query submitted to the web search engine was afterwards submitted to the multimedia vertical. This likelihood was higher for users aged 10 to 19 years old (7% against 4% for the other age groups). We believe that this likelihood is higher than the other events explored due to the functionality provided by state-of-the-art search engines, in which the query submitted to the web vertical is automatically sent to multimedia verticals by clicking on tabs or links pointing to these services. On the other hand the fraction of each one of the browsing events that leads to multimedia search was below 1% for all the age groups.

Conclusion 7: *On average, browsing activities lead to one third of web search events. Slightly higher percentages were observed in the case of adult users. The child user groups were more prone to start the browsing sessions with a search on the web vertical. On the other hand, knowledge browsing events had the highest likelihood to trigger a search, particularly in the youngest group of users. This result points to the importance of providing appropriate educational material that is suitable for young users. The analysis of multimedia usage between age groups pointed to the need for improving access to multimedia results for the youngest group of users.*

7.4 Query recommendations for young users

In Chapter 5, query recommendation methods based on tags from social media and resources for children were explored. The motivation for providing query recommendation for these users comes from the findings presented in Chapter 3. Recall, that the queries of young users were on average shorter, compared to the average web user, and their query vocabulary was significantly smaller. Query recommendation is used to help young users to focus their search on topics that are more likely to be relevant and of interests for them. The two methods proposed in Chapter 5 employed

resources carefully designed for young users and their description with social media, in order to provide high quality query recommendations.

Firstly, a novel biased random walk was proposed to boost keywords that are frequently associated with topics and content designed for young users. The biased random walk is based on information gain metrics that emphasize tags from social media more likely to describe content for children than to describe content for adult users. The method was compared against the random walk proposed by Craswell and Szummer [2007], and three well established biased random walks: *topical page rank*, *spam rank* and the random walk proposed by Fuxman et al. [2008]. All the random walks were trained using the same datasets. Moreover, we contrasted the performance of our methods against the query suggestions of a state-of-the-art search engine. Even though the comparison may not be fair, given that the latter addresses requests for all kinds of public, it is still important to prove that, for young users, our methods provide a significant gain in performance over what the industry provides.

The methods were compared by using two test sets. A large sample of queries and query reformulations submitted by users aged 7 to 18 years old form the Yahoo! Search engine, and a large sample of query reformulations extracted from the AOL logs based on clicks on Dmoz domains, methodology that was used in Chapter 2. The query reformulations were carefully chosen to reflect those clicks that were successful. In this way, the problem of query recommendation can also be seen as a prediction problem, in which the task is to predict the query that led to a successful click.

The evaluation of this method addressed the following research question: *R.Q-5.1 To what extent does a random walk, biased by using information gain metrics, improve the effectiveness of the query recommendations for young users over traditional biased and unbiased random walks?*

In terms of mean average precision, discount cumulative gain *NDCG* and recall, the method proposed clearly outperforms the random walk proposed by Craswell and Szummer [2007] and the three baseline biased random walks. Similarly, a large improvement was observed when contrasting the performance of the query suggestions provided by our method and the query suggestions of the Bing search engine. Two variations of our method were evaluated: using backward and forward probability propagation. Both approaches outperformed the baselines. Nonetheless, the former outperformed the latter. Even though, statistical significant differences were found for the two variations of our method, the differences were marginal in most cases.

Conclusion 8: *The two variations of the biased random walk outperformed state-of-the-art random walks and well established biased random walks using a set of query reformulations submitted by children and teenagers. Large gain margins were observed when contrasting the results of our method against the query recommendation of the Bing search engine. This results shows that modern search engines can greatly benefit*

from the methods presented in this work.

It is important to mention that a larger performance gain was obtained when using the method for the child user segment. Even though, performance gains were also observed for the query reformulations of the teenage users, the margins were smaller than the ones observed for children.

The ranking of the query suggestions provided by the biased random walk was improved using a learning to rank approach. Topic features, language models, seed similarity and query structure features were employed. A extensive evaluation was carried out to understand the gain provided by each feature: *R.Q-5.2. Can we improve the quality of the ranking of query recommendations by combining the random walk score with features based on language models and topical knowledge?*

We found significant improvements in the ranking of results using all the features (measured by *NDCG*). The topic based features were the best performing features along with the language model. The seed similarity feature also led to small performance gain. On the other hand, the string structure features did not have any significant impact on the ranking.

Conclusion 9: *Combining the scores of the random walk proposed with topical features and the language model features derived from the documents oriented to young users led to significant improvements in the quality of the ranking of query suggestions in comparison with state-of-the-art learning to rank methods.*

Gathering a large scale collection of queries submitted by young users can become problematic given the increasing privacy concerns of logging user data. For this reason we were interested in collecting evidence of the feasibility of using queries targeting content for young users (and not necessarily submitted by young users) for the evaluation of query recommendation methods: *R.Q.5-3 Can we substitute successful query reformulations submitted by young users by using query reformulations targeting content for this user segment in the evaluation of query recommendation?*

Consistent results were observed using the query reformulations from the AOL log and the Yahoo! search logs in the evaluation of the biased random walk. The best performing random walk models were aligned with the *kids* and *teens* query sets and with the query reformulations of users aged *10-12* and *13-15*. Further evidence was provided by the high correlation of the topics targeted by both query sets (from AOL and Yahoo! logs), as shown in Chapter 3. These results provide evidence of the suitability of evaluating query reformulations for users aged 10 to 12 years old and 13 to 15 using query reformulations landing on content focused on young users. Nonetheless, further research is required to show that this result holds on different sets of moderated urls (i.e. *KidsClicks*, *Yahoo Kids!*).

Conclusion 10: *Queries targeting content for young users identified through the Dmoz Kids urls are suitable for evaluating methods addressing the problem of query recommendation for children from 10 to 12 years old and teenagers aged 13 to 15 years old.*

7.5 Vertical selection for young users

We hypothesized that young users would benefit from an integrated set of results from suitable search services given the high effort required to access high quality content targeted at young users and their search characteristics, as described in Chapters 2 and 3 respectively.

Chapter 6 envisaged an aggregated search system for young users that integrates search services (i.e. verticals) providing moderated content for children (e.g. Science for Kids), and search services that provide content for all type of public, but that may contain content suitable for young users as well (i.e. YouTube). The settings of the problem are unique in the sense that the search is carried out in a specific information domain (content for children), and that the set of verticals is heterogeneous and may not always provide appropriate content for these users.

Two vertical selection methods were proposed. Recall that vertical selection methods are used to select automatically the most appropriate search services in IR systems given an information request. In this case, we focused on users aged 8 to 12 years old. The first method is based on the estimation of the content that is suitable for children and the content that is available for the average web user in each vertical. The method combines these estimations to measure the likelihood of finding content suitable for children (per vertical). These likelihoods were used in conjunction with a state-of-the-art vertical selection method: *ReDDe*. We refer to the method proposed as *ReDDe-R*. The following research question was addressed to quantify the benefit of this method: *R.Q-6.1. To what extent can we improve state-of-the-art techniques of vertical selection through the estimation of the content available in the verticals for users between 8 to 12 years old?*

The second method of vertical selection employed a language model representation of the verticals based on tags from social media (tags describing url samples representing the verticals) and the query. Social media had not been used before as a resource in the problem of vertical selection. We refer to this method as *LM*. This method was evaluated to address the following research question: *R.Q-6.2. What is the benefit of using tags from social media to represent the query and the verticals in the problem of vertical selection?*

A test collection for the evaluation of vertical selection methods was designed

to address both research questions. The collection utilized a carefully set of queries identified with the methodology described in Chapter 2. A set of verticals was manually selected based on the topic distribution found in Chapter 3. The verticals chosen provide a high coverage of the queries in the collection. The vertical relevant judgement, mapping the verticals that are relevant for each query, were gathered using crowd-sourcing. Two methodologies were used to gather the assessment. Firstly, vertical results were compared to the Google Web results for each query. Assessors were asked to chose between the two types of results and, based on these comparisons, the relevant verticals were derived. In the second approach, the vertical results were evaluated independently, using a graded score. We found that both approaches led to a different set of relevant verticals, particularly the former approach, which was more prone to visual bias, in the sense that image oriented results were preferred by the assessors. Nonetheless, the assessor agreement was substantial for both approaches, thus the two test sets were employed to evaluate the two vertical selection methods proposed.

In the evaluation, two baselines were adopted (*ReDDe* and *Clarity*). We evaluated our two methods (*ReDDe-R* and *LM*) and the combination of *LM* with *ReDDe* (denoted as *LMReDDe*) and *LM* with *ReDDe-R* (denoted as *LMReDDe-R*), which were combined by weighting the scores. The evaluation was carried out by measuring recall, precision and mean average precision *(MAP)*.

We found that the best performing methods for the first test set was *LMReDDe* and *LMReDDe-R*. For the second test set, *LM* was the best performing method, followed by *LMReDDe-R*. The evaluation was also carried out ignoring the results of the Google Web vertical. This experimental setting was motivated by the fact that the Google Web results were not evaluated directly with the first methodology used to gather the assessments, contrary to the second methodology in which the Google Web results are graded. Under these settings, the *LM* was the best performing method using the first test set. For the second test set *LMReDDe-R* led to the best performance. Overall, we observed that *ReDDe-R* outperformed *ReDDe* on most of the experimental settings explored. Similarly, the score of *LM* was utilized in the best performance system for all the experimental settings shown in Chapter 6. These results allow us to answer *R.Q-6.1* and *R.Q-6.2* with the following conclusion:

Conclusion 11: *Using the estimation of the amount of content that is suitable for children in each vertical led to consistent performance improvements in terms of precision and recall in the problem of vertical selection in respect to state-of-the-art methods that do not make use of this information. The usage of tags from social media to represent the user's query and verticals led to the highest performance gain of all the experimental settings employed.*

In the second part of the Chapter 6 a game was presented to compare the interaction of users on three types of result pages: Google Web results without images (i.e. *plain)*), Google Web results (i.e. *google*), and aggregated results from our collection (i.e. *aggregated*). The latter pages were constructed using the relevant assessments with the graded score methodology. In the game, users were asked to select results for a given topic on a result page. Points were awarded based on the number of users that previously clicked on the results of the page. Thirty five queries from the collection were employed. In this study, a group of children from an international primary school in the Netherlands (11 children between 9 and 10 years old) were engaged. A large group of adults was also addressed with crowd-sourcing.

The first aim of the case studies was to evaluate the benefits of presenting results from the verticals of our collection to children in the case of open information tasks. This aim is expressed through the following research question: *R.Q-6.3. Do users aged 8 to 12 years old explore more pages with blended results from the verticals of our collection than pages retrieved from a state-of-the-art search engine? In which type of result pages do they agree more in terms of the content clicked?* The benefits were measured based on the amount of user interaction with each page and the level of agreement.

The game was also employed to compare the vertical preferences of children and adults. The following research question was addressed in this regard: *R.Q-6.4. Which verticals are preferred by children aged 8 to 12 years old given an heterogeneous set of topics, and how do these vertical preferences differ from the preferences of adult users?*

It was found in both case studies (children and adult users) that users interact more with the pages blending results from the verticals of our collection. This interaction was measured in terms of the duration of clicks and the game rounds. Child users were also found to skip to the *aggregated* pages less often than the other page types. Moreover, children agreed more on the results clicked on the *aggregated* pages, which shows that it was easier to identify content that is relevant in the *aggregated* pages. These results allow us to conclude:

Conclusion 12: *In open defined information tasks, both groups of users explored and interacted more in the page results presenting blended results from the verticals of our collection. These pages were preferred over state-of-the-art page results and a simplified version of the results provided by modern search engines. The greater agreement found in the aggregated pages suggested that for children it was easier to find relevant content on these pages.*

In regard to research question *R.Q-6.4*, we measured the agreement between the vertical selection of the group of children and adults through their clicks. We found

that the best agreement was obtained on the *aggregated* interface. However, the agreement was not substantial, which suggests that the benefits in terms of interaction and exploration seen in these types of pages can be improved further by using more accurate vertical relevant judgements. Particularly, we observed that children tended to chose as relevant, results from *Yahoo! Answers* and images, more often than adults. On the other hand, adults selected the *Wikipedia* results more often than the group of children. Overall, high agreement was found on verticals providing moderate and educational content for children, such as *Science for kids* and *KidsHealth*. These verticals also had a high likelihood to be clicked by the group children. From these results we can conclude that:

Conclusion 13: *The agreement between children and adults was fair, in terms of the verticals clicked, although not substantial. The main differences were observed in the preferences of children towards image results and results from the Yahoo! Answers vertical, and the preferences of the adult group towards Wikipedia and the Google Web results. Nonetheless, we showed that even when these differences existed, the aggregated pages were still preferred and capture more interaction and agreement from the group of children.*

7.6 Future Work

The directions in which the work presented in this thesis can be expanded and applied in related domains is summarized in the following paragraphs.

7.6.1 Large Scale Search Behavior of young users

Even though we provided a comprehensive analysis of the search behavior of young users on a large scale, it is still unclear how this behavior varies according the topic that is being searched. For instance, it is reasonable to expect a different type of interaction when young users are searching for school material and for game reviews. Similarly, the impact of the complexity of the information task in terms of search behavior and the search strategies adopted by young users is not clear.

Cultural and other demographic aspects of the population were not explored in this thesis. Understanding the search difficulties of users based on demographics such as educational level, language deficiencies, region or income can lead to solutions of the specific difficulties for this segment of users. For instance, we found, by intersecting the search logs with the US census data, that there is a gap between young users in terms of readability level of clicked results, in specific case of users located

in regions with high and low average income. Note that these findings are not only relevant for the information retrieval community but also for educators and policy makers.

Little work has been carried out to identify the aesthetic characteristics of the websites visited by young users, their impact on the search process and their level of engagement, particularly on a large scale. Features such as font type, size of images and page layout can be explored in detail to understand the key features that are needed to provide engaging content for children [Jochmann-Mannak et al., 2012].

In our analysis we explored the difference in the sentiment expressed in the queries of young and adult users. However, no differences were found. In this regard research is needed to understand the affective factors arising during the search process of young users.

Although not directly related to the case of information retrieval for children, several differences between adults of different ages were found in the analysis carried out in Chapter 3, particularly for the oldest segment of users. The analysis of fine grained groups of senior searchers and the identification of the problems that these users face when they search and consume information on the Internet, represent another important direction for future research.

7.6.2 Query recommendation for young users

Several features can be explored to improve the efficiency of the query recommendation methods proposed in this thesis. Particularly, post-retrieval features have not been explored in this information domain. For instance, the language readability and complexity in the results retrieved can be taken into account in learning to rank the model proposed. Other features derived from the search sessions (e.g. click dwell time) may also provide evidence for better query suggestions that adapt dynamically, according the state of the search session.

Methods for providing multi-media query suggestions represent another interesting line of research. For very young users it may be more useful to not only provide textual query suggestions, but to complement this content with images, audio and even videos. A clear understanding of the types of search tasks in which it is convenient to complement the query suggestions with rich media is still needed.

7.6.3 Aggregated search for young users

More sophisticated methods for combining the scores of the vertical selection methods proposed in this thesis can lead to better performance. For instance, learning to rank the approach utilized for query recommendation can be utilized in the context

of vertical selection. Moreover, as it was mentioned before, the readability and sentiment characteristics of the verticals' documents are some of the features that can be explored to improve the quality of the vertical selection. Features derived from the search session can also be employed to select verticals dynamically during the search process. Little research has been dedicated to account for the coherence between the vertical results, which have been shown to have an impact in the interaction with the result page [Arguello and Capra, 2012]. Similarly, the redundancy introduced by the results of different verticals have not been addressed in depth in the literature.

Further research is also required to improve the benefits of presenting aggregated pages to young users. For instance, we did not explore the effect of varying the order in which the vertical results are presented and the ideal number of vertical results to present given the characteristics of the information task. Evidence of the benefits in terms of interaction was provided for open information tasks. However, research to identify the benefits of this type of result page for close information tasks and complex, multi-session tasks, need to be investigated. For instance, aggregated interfaces may help children to not only solve complex informational tasks, but also to retain more information (e.g. school content).

It was shown that groups of children and adults do not always agree in the selection of verticals. We believe that better aggregated pages can be built by either improving the agreement between the two group of users or by reaching a larger audience of children. This can be achieved through the design of engaging games aimed at capturing their result preferences. An example of this type of games has been provided in this thesis. For instance, this game can be generalized to allow children to submit queries in order to study multi-session information tasks. Cooperative scenarios in which children help each other can also be explored by allowing children to team up in order to solve information tasks of different complexity.

The agreement between the group of users can be improved by exploring other methodologies to gather assessments from adult users, for instance by allowing only parents to submit vertical judgements. Also, other methods of reducing the bias introduced by adult judges are worth exploring. We believe this problem is relevant given the current difficulty in the gathering, on a large scale, of the relevant judgements for children.

7.7 Final Remarks

In this thesis we provide a comprehensive analysis of search difficulties experienced by children along with their search and browsing behavior, which has not been done with commercial search engines on a large scale before. The results of our study have important implications for the industry, educators and the research community.

Concrete recommendations are given to improve the search success of young users in state-of-the-art search engines.

We focus on two key solutions: Query recommendations and vertical selection for young users. We show, in this thesis, that the query recommendation methods proposed provide significant improvements over state-of-the-art methods and over what the industry currently offers to young users. We believe that the methods proposed can greatly benefit modern search engines in helping young users in the formulation of the search queries, which is one of their main difficulties when searching, as identified in this thesis. The vertical selection methods proposed were shown to out-perform previous methods in the domain of information for young users. Children prefer the aggregated result pages with results from verticals showing moderated content and they explore and interact more with the content displayed on these pages. The lack of exploration of the result pages is also one of the prominent problems that young users face during the search process. This solution has great potential in an educational context, allowing young users to search high quality educational content in an engaging and transparent manner.

Finally, we believe that this thesis provides a solid basis for providing more sophisticated methods of understanding and supporting the increasing number of young users online. They are set to play an increasingly large role on the Internet, given that every day users go online at earlier ages.

Appendix A

Macro-averaged results for the AOL search logs

In the following paragraphs we report macro-averaged results for the log analysis carried out in Chapter 2. Macro-averaged results are obtained by averaging each metric (e.g. query length) per user and then averaging across users. As it was the case with the results reported in Chapter 2, most of the results were statistical significant using the paired t-test at 0.1% confidence level (please refer to Section 2.5 for more details about the statistical tests). Each non-statistical significant result is reported in this appendix.

Figure A.1 presents the macro-averaged query frequency distribution. The equivalent micro-averaged distribution is shown in Figure 2.6. Non-statistical results were found between the *kids* and *teens* query set at length 4, and between the *kids*, *teens* and *mteens* at 8,9 at 10.

Similarly the macro-averaged rank distribution is displayed in Figure A.2. The micro-averaged version of the rank distribution is shown in Figure 2.5. Non-statistical results were found at rank position 2 between the *teens* and *mteens* sets and at 10,11,12,13,14 between the *kids*, *teens* and *mteens* sets.

Figures A.3 and A.4 shows the macro-average session length (in number of entries) and session duration distribution (in minutes) respectively. The micro-averaged distribution are shown in 2.7 and 2.8 respectively. All the results were statistical significant except at 6 for *kids* and *teens* and at 7,8,9 and 10 between the *kids*, *teens*, *mteens* sets.

Table A.1 presents the macro-averaged session length and session duration for all the query sets. Equivalent micro-averaged results are shown in Table 2.7.

Table A.2 present the macro-averaged query reformulations defined in Section

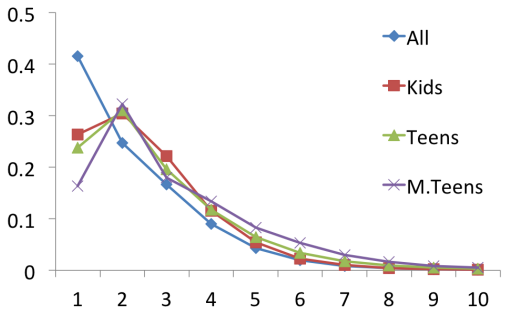


Figure A.1: Q. length distribution (macro)

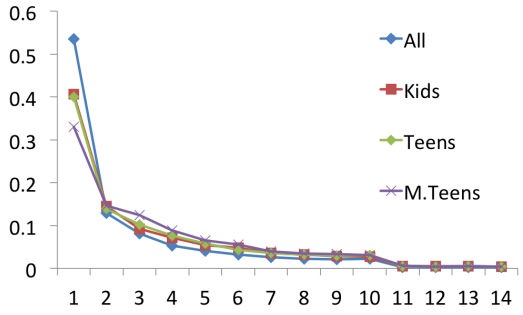


Figure A.2: Rank distribution (macro)

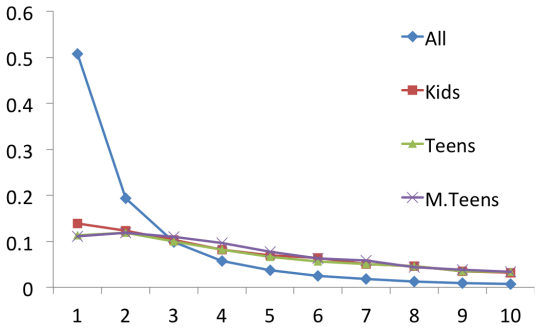


Figure A.3: Sessions length distribution (macro)

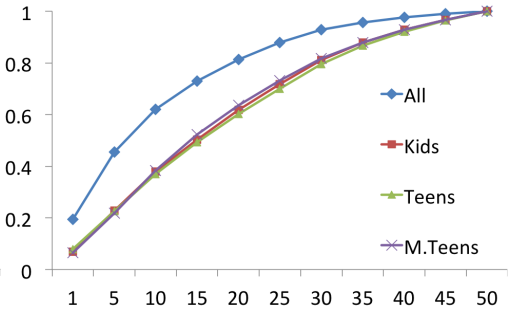


Figure A.4: Session duration distribution (macro)

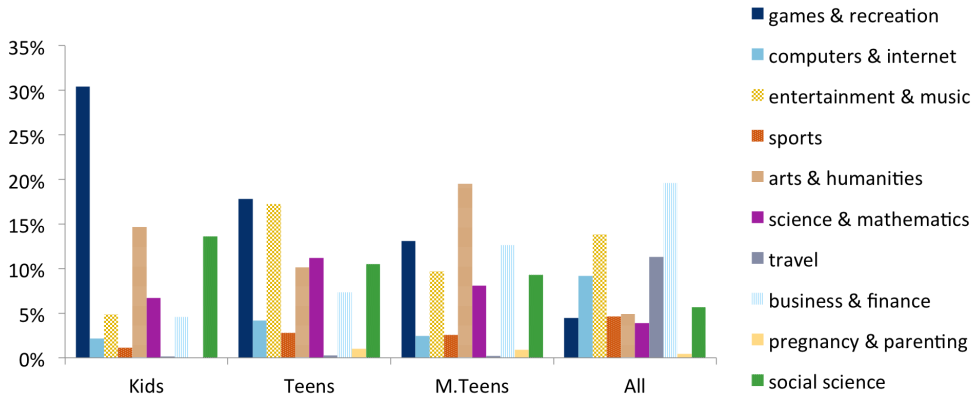


Figure A.5: Macro topic distribution of each data set (Yahoo! Directory categories)

	All	Kids	Teens	M.Teens
S. length (mean)	2.8	8.6	10.0	8.8
S. duration (mean)	5.6	10.8	14.0	12.9

Table A.1: Session average length and duration (macro)

	All	Kids	Teens	M.Teens
n.q	67.4%	43.0%	43.2%	41.1%
w.a.	1.0%	1.4%	1.6%	1.9%
w.r.	0.2%	0.3%	0.4%	0.3%
w.c.	3.7%	6.8%	6.6%	7.5%
m.r.	24.8%	45.3%	44.4%	45.4%
p.q.	1.5%	1.5%	1.6%	1.4%
s.c.	1.4%	1.7%	2.2%	2.3%

Table A.2: Query reformulation proportions (macro)

2.7.3. Micro-averaged results are shown in Table 2.8. The results displayed are normalized for each query set (each column summing up to 1). The following were the non-statistical results at 0.1% with the paired t-test: *w.r* and *p.q.* between all the sets and *s.c.* between *teens* and *mteens*.

Table A.3 presents the results for the click patterns for each one of the query reformulation types. In this case each row sum ups to 1. Micro-averaged results are shown in Table 2.9. The following results were not found statistical significant: For the query reformulation *n.q* the patterns *Click-Skip* and *Skip-Skip* between the *kids*, *teens* and *mteens* sets. For *w.a.*, *w.r.* and *w.c* and the the pattern *Skip-Click* for the comparisons between all data sets. For *w.c* and the pattern *Click-Skip* between the *kids* and *teens* sets. For *m.r.* and the patterns *Click-Skip* and *Skip-Click* for the comparisons between the *kids*, *teens* and *mteens* sets. For *p.q.* and the pattern *Skip-Skip* for the three sets and finally *s.c.* and the pattern *Skip-Click* for the comparison between the *kids* and *teens* sets.

Figure A.5 presents the macro topic distribution of the set of informational queries identified in the AOL search log. We displayed the topic distribution for the *kids*, *teens*, *mteens* and from the sample of queries from the whole data set *average user*. Figure 2.4 shows the micro-averaged results.

	Dataset	Click-Click	Click-Skip	Skip-Click	Skip-Skip	Mean rank change	Mean time average
n.q	kids	36.5%	20.1%	25.6%	17.8%	-2.3	402.1
	teens	36.2%	20.4%	26.0%	17.4%	-2.1	394.4
	m.teens	38.5%	20.9%	23.5%	17.1%	-2.0	421.9
	all	15.3%	13.9%	22.5%	48.3%	-0.5	311.7
w.a.	kids	31.4%	13.3%	30.5%	24.9%	-4.2	164.4
	teens	35.4%	11.7%	29.7%	23.1%	-3.1	175.7
	m.teens	30.9%	14.8%	31.8%	22.5%	-3.4	174.9
	all	16.6%	11.3%	29.7%	42.5%	-1.5	150.1
w.r.	kids	25.9%	15.2%	30.2%	28.6%	-5.8	180.2
	teens	29.4%	6.2%	30.1%	34.3%	-3.9	152.0
	m.teens	27.9%	13.7%	30.6%	27.8%	-4.8	164.2
	all	15.1%	10.0%	30.6%	44.3%	-1.8	156.7
w.c.	kids	44.0%	14.6%	24.6%	16.8%	-4.1	278.2
	teens	42.7%	14.8%	24.0%	18.6%	-4.2	264.8
	m.teens	43.1%	16.2%	24.5%	16.1%	-3.1	287.7
	all	23.6%	13.5%	24.7%	38.2%	-1.3	248.8
m.r.	kids	78.2%	4.7%	6.0%	11.2%	1.5	57.4
	teens	76.8%	4.5%	6.0%	12.6%	1.4	54.8
	m.teens	79.1%	4.6%	5.9%	10.4%	1.6	54.3
	all	46.7%	4.9%	7.7%	40.7%	0.8	88.2
p.q.	kids	23.4%	18.5%	27.9%	30.2%	-0.5	236.2
	teens	25.0%	18.8%	26.0%	30.1%	-0.4	242.2
	m.teens	21.7%	19.7%	28.2%	30.4%	0.1	234.8
	all	12.7%	15.6%	21.7%	50.0%	-0.1	214.2
s.c.	kids	9.2%	4.4%	56.5%	29.9%	-1.3	71.9
	teens	10.2%	5.9%	56.4%	27.5%	-1.0	77.1
	m.teens	11.0%	4.5%	57.9%	26.6%	-1.1	73.7
	all	5.3%	3.6%	43.6%	47.6%	-0.4	77.6

Table A.3: Click pattern frequency for the query reformulation types (macro)

Appendix B

Case study survey

UNIVERSITEIT TWENTE.

Faculteit Elektrotechniek, Wiskunde en Informatica

Databases Group. Carried out by Sergio Duarte and Dr. Djoerd Hiemstra.

Survey: Game session follow up

Hero name: .....

Did you find the game hard?

.....

.....

What did you like and did not like about the game?

.....

.....

.....

Did you notice differences in the pages displayed in each game session?

.....

.....

.....

If yes, Did you find the game more fun, or easier, for a specific type of page?

.....

.....

.....

Did you notice that sometimes results are special and group together in blocks? For example images, YouTube video. If yes, which ones.

.....

.....

.....

Did you find any of these blocks of special results more useful?

.....

.....

What do you expect if you can click and go to the special results?

.....

.....

List of Publications

Sergio Duarte Torres. Information retrieval for children based on the aggregated search paradigm. Technical Report TR-CTIT-11-05, Centre for Telematics and Information Technology University of Twente, Enschede, March 2011. 7, 112

Sergio Duarte Torres and Ingmar Weber. What and how children search on the web. In *CIKM '11*, pages 393–402, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0717-8. doi: 10.1145/2063576.2063638. URL `http://doi.acm.org/10.1145/2063576.2063638`. 11, 45, 124, 131

Sergio Duarte Torres, Djoerd Hiemstra, and Pavel Serdyukov. An analysis of queries intended to search information for children. In *Proceedings of the third symposium on Information interaction in context*, IIiX '10, pages 235–244, New York, NY, USA, 2010a. ACM. ISBN 978-1-4503-0247-0. doi: http://doi.acm.org/10.1145/1840784.1840819. URL `http://doi.acm.org/10.1145/1840784.1840819`. 10, 13

Sergio Duarte Torres, Djoerd Hiemstra, and Pavel Serdyukov. Query log analysis in the context of information retrieval for children. In *SIGIR '10*, pages 847–848, New York, NY, USA, 2010b. ACM. 10, 13

Sergio Duarte Torres, Djoerd Hiemstra, Ingmar Weber, and Pavel Serdyukov. Query recommendation for children. In *CIKM '12*, pages 2010–2014, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1156-4. doi: 10.1145/2396761.2398562. URL `http://doi.acm.org/10.1145/2396761.2398562`. 11, 107

Sergio Duarte Torres, Djoerd Hiemstra, and Theo Huibers. Vertical selection in the information domain of children. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '13, pages 57–66, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2077-1. doi: 10.1145/2467696.2467714. URL `http://doi.acm.org/10.1145/2467696.2467714`. 11, 143

Sergio Duarte Torres, Djoerd Hiemstra, and Ingmar Weber. Analysis of search and browsing behavior of children on the web. (To be published). *Transactions on the Web (TWEB)*, 8, 2014a. 11, 87

Sergio Duarte Torres, Djoerd Hiemstra, Ingmar Weber, and Pavel Serdyukov. Query recommendation in the domain of information for children. (To be published). *Journal of the American Society for Information Science and Technology (JASIST)*, 2014b. 11, 107

Frans van der Sluis, Sergio Duarte Torres, Djoerd Hiemstra, Betsy van Dijk, and Frea Kruisinga. Visual exploration of health information for children. In *ECIR'11*, pages 788–792, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-20160-8. URL `http://dl.acm.org/citation.cfm?id=1996889.1997004`. 110

Leif Azzopardi, Doug Dowie, Sergio Duarte, Carsten Eickhoff, Richard Glassey, Karl Gyllstrom, Djoerd Hiemstra, Franciska de Jong, Frea Kruisinga, Kelly Marshall, Sien Moens, Tamara Polajnar, Frans van der Sluis, and Arjen de Vries. Emse: supporting children's information needs within a hospital environment. In *ECIR '12*, pages 578–580, Berlin, Heidelberg, 2012a. Springer-Verlag. ISBN 978-3-642-28996-5. doi: 10.1007/978-3-642-28997-2_70. URL `http://dx.doi.org/10.1007/978-3-642-28997-2_70`. 99, 110

References

Tony Abou-Assaleh, Tapajyoti Das, Weizheng Gao, Yingbo Miao, Philip O'Brien, and Zhen Zhen. A link-based ranking scheme for focused search. In *WWW '07*, pages 1125–1126, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7. doi: 10.1145/1242572.1242727. URL `http://doi.acm.org/10.1145/1242572.1242727`. 111

Eugene Agichtein, Eric Brill, and Susan Dumais. Improving web search ranking by incorporating user behavior information. In *SIGIR '06*, pages 19–26, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7. doi: 10.1145/1148170.1148177. URL `http://doi.acm.org/10.1145/1148170.1148177`. 195

Soontae An and Susannah Stern. Mitigating the effects of advergames on children; do advertising breaks work? *Journal of Advertising*, 40(1):43 – 56, 2011. 60

Jaime Arguello and Robert Capra. The effect of aggregated search coherence on search behavior. In *CIKM '12*, pages 1293–1302, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1156-4. doi: 10.1145/2396761.2398432. URL `http://doi.acm.org/10.1145/2396761.2398432`. 148, 211

Jaime Arguello, Fernando Diaz, Jamie Callan, and Jean-Francois Crespo. Sources of evidence for vertical selection. In *SIGIR '09*, pages 315–322, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi: http://doi.acm.org/10.1145/1571941.1571997. 44, 144, 147, 161, 162

Jaime Arguello, Fernando Diaz, and Jean-François Paiement. Vertical selection in the presence of unlabeled verticals. In *SIGIR '10*, pages 691–698, 2010. 144, 147, 151, 161

Jaime Arguello, Fernando Diaz, Jamie Callan, and Ben Carterette. A methodology for evaluating aggregated search results. In *ECIR '11*, pages 141–152, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-20160-8. URL `http://dl.acm.org/citation.cfm?id=1996889.1996909`. 147, 153, 174

Jaime Arguello, Wan-Ching Wu, Diane Kelly, and Ashlee Edwards. Task complexity, vertical display and user interaction in aggregated search. In *SIGIR '12*, pages 435–444, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1472-5. doi: 10.1145/2348283.2348343. URL `http://doi.acm.org/10.1145/2348283.2348343`. 148, 155

Leif Azzopardi, Doug Dowie, Sergio Duarte, Carsten Eickhoff, Richard Glassey, Karl Gyllstrom, Djoerd Hiemstra, Franciska de Jong, Frea Kruisinga, Kelly Marshall, Sien Moens, Tamara Polajnar, Frans van der Sluis, and Arjen de Vries. Emse: supporting children's information needs within a hospital environment. In *ECIR '12*, pages 578–580, Berlin, Heidelberg, 2012a. Springer-Verlag. ISBN 978-3-642-28996-5. doi: 10.1007/978-3-642-28997-2_70. URL `http://dx.doi.org/10.1007/978-3-642-28997-2_70`. 99, 110

Leif Azzopardi, Douglas Dowie, and Kelly Ann Marshall. Yoosee: a video browsing application for young children. In *SIGIR '12*, pages 1017–1017, New York, NY, USA, 2012b. ACM. ISBN 978-1-4503-1472-5. doi: 10.1145/2348283.2348442. URL `http://doi.acm.org/10.1145/2348283.2348442`. 110

Leif Azzopardi, Douglas Dowie, Kelly Ann Marshall, and Richard Glassey. Mase: create your own mash-up search interface. In *SIGIR '12*, pages 1008–1008, New York, NY, USA, 2012c. ACM. ISBN 978-1-4503-1472-5. doi: 10.1145/2348283.2348433. URL `http://doi.acm.org/10.1145/2348283.2348433`. 110

Ricardo Baeza-Yates and Alessandro Tiberi. Extracting semantic relations from query logs. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 76–85, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-609-7. doi: http://doi.acm.org/10.1145/1281192.1281204. URL `http://doi.acm.org/10.1145/1281192.1281204`. 107, 109

Ricardo Baeza-Yates, Liliana Calderón-Benavides, and Cristina González-Caro. The intention behind web queries. pages 98–109. 2006. doi: 10.1007/11880561_9. URL `http://dx.doi.org/10.1007/11880561_9`. 13

Jamshid Beheshti, Andrew Large, and Marni Tam. Transaction logs and search patterns on a children's portal/journaux de transaction et modes de recherche sur un portail web destiné aux enfants. *Canadian Journal of Information and Library Science*, 34(4): 391–402, 2010. 50

N. J. Belkin, C. Cool, D. Kelly, G. Kim, J. y. Kim, H. j. Lee, G. Muresan, M. c. Tang, and X. j. Yuan. Query length in interactive information retrieval. In *SIGIR '03*, pages 205–212, New York, NY, USA, 2003. ACM. 23

Michael Bendersky and W. Bruce Croft. Analysis of long queries in a large scale search log. In *WSCD '09: Proceedings of the 2009 workshop on Web Search Click Data*, pages 8–14, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-434-8. doi: http://doi.acm.org/10.1145/1507509.1507511. 23

Adam Bermingham and Alan F. Smeaton. A study of inter-annotator agreement for opinion retrieval. In *SIGIR '09*, pages 784–785, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1572127. URL `http://doi.acm.org/10.1145/1571941.1572127`. 159

Dania Bilal. Children's use of the yahooligans! web search engine: I. cognitive, physical, and affective behaviors on fact-based search tasks. *JASIS*, 51(7):646–665, 2000. 16, 18, 40, 42, 46, 47, 50

Dania Bilal. Children's use of the yahooligans! web search engine ii. cognitive and physical behaviors on research tasks. *J. Am. Soc. Inf. Sci. Technol.*, 52(2):118–136, 2001. ISSN 1532-2882. doi: http://dx.doi.org/10.1002/1097-4571(2000)9999:9999⟨::AID-ASI1038⟩3.3.CO;2-I. 2, 6, 9, 40, 42, 46

Dania Bilal. Children's use of the yahooligans! web search engine iii. cognitive and physical behaviors on fully self-generated search tasks. *J. Am. Soc. Inf. Sci. Technol.*, 53(13): 1170–1183, 2002. ISSN 1532-2882. doi: http://dx.doi.org/10.1002/asi.10145. 2, 6, 16, 18, 21, 24, 36, 41, 42, 46, 47, 50, 84, 122

Dania Bilal and Jinx Stapelton Watson. Children's paperless projects: Inspiring research via the web. In *IFLA. Conference*, 1998. 1, 16, 57

Lidong Bing, Wai Lam, and Tak-Lam Wong. Using query log and social tagging to refine queries based on latent topics. In *CIKM '11*, pages 583–592, 2011. 110, 123

Paolo Boldi, Francesco Bonchi, Carlos Castillo, Debora Donato, Aristides Gionis, and Sebastiano Vigna. The query-flow graph: model and applications. In *CIKM '08*, pages 609–618, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. doi: http://doi.acm.org/10.1145/1458082.1458163. 13, 19, 21

Paolo Boldi, Francesco Bonchi, Carlos Castillo, Debora Donato, and Sebastiano Vigna. Query suggestions using query-flow graphs. In *Proceedings of the 2009 workshop on Web Search Click Data*, WSCD '09, pages 56–63, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-434-8. doi: 10.1145/1507509.1507518. URL `http://doi.acm.org/10.1145/1507509.1507518`. 109, 112, 114

Christine L. Borgman, Sandra G. Hirsh, Virginia A. Walter, and Andrea L. Gallagher. Children's searching behavior on browsing and keyword online catalogs: the science library catalog project. *JASIS*, 46(9):663–684, 1995. 15

Leanne Bowler, Andrew Large, and Gill Rejskind. Primary school students, information literacy and the web. *Education for Information*, 19(3):201–224, 2001. 17

David J. Brenes and Daniel Gayo-Avello. Stratified analysis of aol query log. *Inf. Sci.*, 179 (12):1844–1858, 2009. ISSN 0020-0255. doi: http://dx.doi.org/10.1016/j.ins.2009.01.027. 19, 36

Elana Broch. Children's search engines from an information search process perspective. *School Library Media Research*, 3, 2000. 1, 15, 16

Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002. ISSN 0163-5840. URL http://www.acm.org/sigs/sigir/forum/F2002/broder.pdf. 13, 26, 27, 53

Marc Bron, Jasmijn van Gorp, Frank Nack, Lotte Belice Baltussen, and Maarten de Rijke. Aggregated search interface preferences in multi-session search tasks. In *SIGIR '13*, pages 123–132, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2034-4. doi: 10.1145/2484028.2484050. URL http://doi.acm.org/10.1145/2484028.2484050. 149

Mary C Burton and Joseph B Walther. The value of web log data in use-based design and testing. *Journal of Computer-Mediated Communication*, 6(3):0–0, 2001. 20

Gianfranco Cecchin. Hypothesizing, circularity, and neutrality revisited: An invitation to curiosity. *Family process*, 26(4):405–413, 1987. 25, 55

Sergiu Chelaru, Ismail Sengor Altingovde, and Stefan Siersdorfer. Analyzing the polarity of opinionated queries. In *Proceedings of the 34th European conference on Advances in Information Retrieval*, ECIR '12, pages 463–467, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-28996-5. doi: 10.1007/978-3-642-28997-2_42. URL http://dx.doi.org/10.1007/978-3-642-28997-2_42. 76

Zhicong Cheng, Bin Gao, and Tie-Yan Liu. Actively predicting diverse search intent from user browsing behaviors. In *WWW '10*, pages 221–230, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772714. URL http://doi.acm.org/10.1145/1772690.1772714. 2, 87, 90, 92, 96, 103

Steve Chien and Nicole Immorlica. Semantic similarity between search engine queries using temporal correlation. In *WWW '05*, pages 2–11, New York, NY, USA, 2005. ACM. ISBN 1-59593-046-9. doi: http://doi.acm.org/10.1145/1060745.1060752. 13

Child Trends Data Bank. Home computer access and Internet use, 2013. 1

Kenneth Church. How many multiword expressions do people know? In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, MWE '11, pages 137–144, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-97-8. URL http://dl.acm.org/citation.cfm?id=2021121.2021152. 30

Kevyn Collins-Thompson, Paul N Bennett, Ryen W White, Sebastian de la Chica, and David Sontag. Personalizing web search results by reading level. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 403–412. ACM, 2011. 77

Nick Craswell and Martin Szummer. Random walks on the click graph. In *SIGIR '07*, pages 239–246, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277784. URL http://doi.acm.org/10.1145/1277741.1277784. 109, 113, 123, 204

Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *WSDM '08*, pages 87–94, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-927-2. doi: 10.1145/1341531.1341545. URL `http://doi.acm.org/10.1145/1341531.1341545`. 58, 124

W. Bruce Croft. Effective text retrieval based on combining evidence from the corpus and users. *IEEE Expert*, 10(6):59–63, 1995. 8

Doug Downey, Susan Dumais, Dan Liebling, and Eric Horvitz. Understanding the relationship between searchers' queries and information goals. In *CIKM '08*, pages 449–458, New York, NY, USA, 2008. ACM. 23, 150

Allison Druin, Elizabeth Foss, Leshell Hatley, Evan Golub, Mona Leigh Guha, Jerry Fails, and Hilary Hutchinson. How children search the internet with keyword interfaces. In *IDC '09: Proceedings of the 8th International Conference on Interaction Design and Children*, pages 89–96, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-395-2. doi: http://doi.acm.org/10.1145/1551788.1551804. 1, 17, 24, 25, 42, 46, 56, 60, 61, 84, 88, 107, 122, 170

Allison Druin, Elizabeth Foss, Hilary Hutchinson, Evan Golub, and Leshell Hatley. Children's roles using keyword search interfaces at home. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 413–422, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-929-9. doi: http://doi.acm.org/10.1145/1753326.1753388. URL `http://doi.acm.org/10.1145/1753326.1753388`. 1, 9, 17, 96, 101

Sergio Duarte Torres. Information retrieval for children based on the aggregated search paradigm. Technical Report TR-CTIT-11-05, Centre for Telematics and Information Technology University of Twente, Enschede, March 2011. 7, 112

Sergio Duarte Torres and Ingmar Weber. What and how children search on the web. In *CIKM '11*, pages 393–402, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0717-8. doi: 10.1145/2063576.2063638. URL `http://doi.acm.org/10.1145/2063576.2063638`. 11, 45, 124, 131

Sergio Duarte Torres, Djoerd Hiemstra, and Pavel Serdyukov. An analysis of queries intended to search information for children. In *Proceedings of the third symposium on Information interaction in context*, IIiX '10, pages 235–244, New York, NY, USA, 2010a. ACM. ISBN 978-1-4503-0247-0. doi: http://doi.acm.org/10.1145/1840784.1840819. URL `http://doi.acm.org/10.1145/1840784.1840819`. 10, 13

Sergio Duarte Torres, Djoerd Hiemstra, and Pavel Serdyukov. Query log analysis in the context of information retrieval for children. In *SIGIR '10*, pages 847–848, New York, NY, USA, 2010b. ACM. 10, 13

Sergio Duarte Torres, Djoerd Hiemstra, Ingmar Weber, and Pavel Serdyukov. Query recommendation for children. In *CIKM '12*, pages 2010–2014, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1156-4. doi: 10.1145/2396761.2398562. URL `http://doi.acm.org/10.1145/2396761.2398562`. 11, 107

Sergio Duarte Torres, Djoerd Hiemstra, and Theo Huibers. Vertical selection in the information domain of children. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '13, pages 57–66, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2077-1. doi: 10.1145/2467696.2467714. URL `http://doi.acm.org/10.1145/2467696.2467714`. 11, 143

Sergio Duarte Torres, Djoerd Hiemstra, and Ingmar Weber. Analysis of search and browsing behavior of children on the web. (To be published). *Transactions on the Web (TWEB)*, 8, 2014a. 11, 87

Sergio Duarte Torres, Djoerd Hiemstra, Ingmar Weber, and Pavel Serdyukov. Query recommendation in the domain of information for children. (To be published). *Journal of the American Society for Information Science and Technology (JASIST)*, 2014b. 11, 107

Jacquelynne Eccles, Allan Wigfield, Rena D. Harold, and Phyllis Blumenfeld. Age and gender differences in children's self- and task perceptions during elementary school. *Child Development*, 64(3):830–847, 1993. 65

Jacquelynne S. Eccles, Janis E. Jacobs, and Rena D. Harold. Gender role stereotypes, expectancy effects, and parents' socialization of gender differences. *Journal of Social Issues*, 46(2):183–201, 1990. 65

Carsten Eickhoff, Pavel Serdyukov, and Arjen P. de Vries. Web page classification on child suitability. In *CIKM '10*, pages 1425–1428, 2010. 110, 120

Carsten Eickhoff, Pavel Serdyukov, and Arjen P. de Vries. A combined topical/non-topical approach to identifying web sites for children. In *WSDM '11*, pages 505–514, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0493-1. doi: 10.1145/1935826.1935900. URL `http://doi.acm.org/10.1145/1935826.1935900`. 149

Carsten Eickhoff, Christopher G. Harris, Arjen P. de Vries, and Padmini Srinivasan. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *SIGIR '12*, pages 871–880, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1472-5. doi: 10.1145/2348283.2348400. URL `http://doi.acm.org/10.1145/2348283.2348400`. 171

Songhe Feng, Congyan Lang, and Bing Li. Towards relevance and saliency ranking of image tags. In *Proceedings of the 20th ACM international conference on Multimedia*, MM '12, pages 917–920, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1089-5. doi: 10.1145/2393347.2396346. URL `http://doi.acm.org/10.1145/2393347.2396346`. 111

Raya Fidel, Rachel K Davies, Mary H Douglass, Jenny K Holder, Carla J Hopkins, Elisabeth J Kushner, Bryan K Miyagishima, and Christina D Toney. A visit to the information mall: Web searching behavior of high school students. *Journal of the American Society for Information Science*, 50(1):24–37, 1999. 17, 50, 58

Elizabeth Foss, Allison Druin, Robin Brewer, Phillip Lo, Luis Sanchez, Evan Golub, and Hilary Hutchinson. Children's search roles at home: Implications for designers, researchers, educators, and parents. *JASIST*, 63(3):558–573, 2012. 6, 151

Ariel Fuxman, Panayiotis Tsaparas, Kannan Achan, and Rakesh Agrawal. Using the wisdom of the crowds for keyword generation. In *WWW '08*, pages 61–70, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367506. URL `http://doi.acm.org/10.1145/1367497.1367506`. 111, 112, 117, 118, 119, 125, 204

Wei Gao, Cheng Niu, Jian-Yun Nie, Ming Zhou, Kam-Fai Wong, and Hsiao-Wuen Hon. Exploiting query logs for cross-lingual query suggestions. *ACM Trans. Inf. Syst.*, 28:6:1–6:33, June 2010. ISSN 1046-8188. doi: http://doi.acm.org/10.1145/1740592.1740594. URL `http://doi.acm.org/10.1145/1740592.1740594`. 109

Reynaldo Gil-García and Aurora Pons-Porrata. Hierarchical star clustering algorithm for dynamic document collections. In *CIARP '08: Proceedings of the 13th Iberoamerican congress on Pattern Recognition*, pages 187–194, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-85919-2. doi: http://dx.doi.org/10.1007/978-3-540-85920-8_23. 28

Sharad Goel, Jake M. Hofman, and M. Irmak Sirer. Who does what on the web: A large-scale study of browsing behavior. In *ICWSM*, Toronto, 2012. AAAI. 2, 90

Tatiana Gossen, Thomas Low, and Andreas Nürnberger. What are the real differences of children's and adults' web search. In *SIGIR '11*, pages 1115–1116, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0757-4. doi: 10.1145/2009916.2010076. URL `http://doi.acm.org/10.1145/2009916.2010076`. 50

Holly Gunn and Gary Hepburn. Seeking information for school purposes on the internet. *Canadian Journal of Learning and Technology/La revue canadienne de l'apprentissage et de la technologie*, 29(1), 2003. 17

Sheng Guo and Naren Ramakrishnan. Mining linguistic cues for query expansion: applications to drug interaction search. In *CIKM '09*, pages 335–344, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-512-3. doi: http://doi.acm.org/10.1145/1645953.1645998. 28

K. Gyllstrom and M.F Moens. A picture is worth a thousand search results: Finding child-oriented multimedia results with collage. In *SIGIR '10*. 7

Karl Gyllstrom and Marie-Francine Moens. Wisdom of the ages: toward delivering the children's web with the link-based agerank algorithm. In *CIKM '10*, pages 159–168, New

York, NY, USA, 2010. ACM. ISBN 978-1-4503-0099-5. doi: http://doi.acm.org/10.1145/1871437.1871462. URL http://doi.acm.org/10.1145/1871437.1871462. 110, 120

Karl Gyllstrom and Marie-Francine Moens. Clash of the typings: finding controversies and children's topics within queries. In *ECIR '11*, pages 80–91, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-20160-8. URL http://dl.acm.org/citation.cfm?id=1996889.1996903. 110

Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. Combating web spam with trustrank. In *VLDB '04*, pages 576–587. VLDB Endowment, 2004. ISBN 0-12-088469-0. URL http://dl.acm.org/citation.cfm?id=1316689.1316740. 111

Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. Beyond dcg: user behavior as a predictor of a successful search. In *WSDM '10*, pages 221–230, New York, NY, USA, 2010a. ACM. ISBN 978-1-60558-889-6. doi: 10.1145/1718487.1718515. URL http://doi.acm.org/10.1145/1718487.1718515. 150

Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. Beyond DCG: user behavior as a predictor of a successful search. In *WSDM '10*, pages 221–230, New York, NY, USA, 2010b. ACM. ISBN 978-1-60558-889-6. doi: http://doi.acm.org/10.1145/1718487.1718515. URL http://doi.acm.org/10.1145/1718487.1718515. 59, 124

Taher H. Haveliwala. Topic-sensitive pagerank. In *WWW '02*, pages 517–526, New York, NY, USA, 2002. ACM. ISBN 1-58113-449-5. doi: 10.1145/511446.511513. URL http://doi.acm.org/10.1145/511446.511513. 111, 116, 117, 125

Taher H Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *Knowledge and Data Engineering, IEEE Transactions on*, 15(4):784–796, 2003. 111

Daqing He and Ayse Goker. Detecting session boundaries from web user logs. In *In Proceedings of the BCS-IRSG 22nd Annual Colloquium on Information Retrieval Research*, pages 57–66, 2000. 19, 21

Andrea B. Hellman. Vocabulary size and depth of word knowledge in adult-onset second language acquisition. *International Journal of Applied Linguistics*, 21(2):162–182, 2011. ISSN 1473-4192. doi: 10.1111/j.1473-4192.2010.00265.x. URL http://dx.doi.org/10.1111/j.1473-4192.2010.00265.x. 30

Djoerd Hiemstra. Term-specific smoothing for the language modeling approach to information retrieval: the importance of a query term. In *SIGIR '02*, pages 35–41, 2002. 162, 163

Djoerd Hiemstra, Claudia Hauff, Franciska de Jong, and Wessel Kraaij. Sigir's 30th anniversary: an analysis of trends in ir research and the topology of its community. *SIGIR Forum*, 41(2):18–24, December 2007. ISSN 0163-5840. doi: 10.1145/1328964.1328966. URL http://doi.acm.org/10.1145/1328964.1328966. 69

Nadine Höchstötter and Dirk Lewandowski. What users see - structures in search engine results pages. *Inf. Sci.*, 179(12):1796–1812, May 2009. ISSN 0020-0255. doi: 10.1016/j.ins.2009.01.028. URL http://dx.doi.org/10.1016/j.ins.2009.01.028. 58

Guichun Hua, Min Zhang, Yiqun Liu, Shaoping Ma, and Liyun Ru. Automatically generating labels based on unified click model. In *WWW '11*, pages 59–60, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0637-9. doi: 10.1145/1963192.1963223. URL http://doi.acm.org/10.1145/1963192.1963223. 124

Jeff Huang and Efthimis N. Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In *CIKM '09*, pages 77–86, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-512-3. doi: http://doi.acm.org/10.1145/1645953.1645966. 37, 38, 122

Jeff Huang, Thomas Lin, and Ryen W. White. No search result left behind: branching behavior with browser tabs. In *WSDM '12*, pages 203–212, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-0747-5. doi: 10.1145/2124295.2124322. URL http://doi.acm.org/10.1145/2124295.2124322. 92

Kai Hui, Bin Gao, Ben He, and Tie-jian Luo. Sponsored search ad selection by keyword structure analysis. In *ECIR '13*, pages 230–241, Berlin, Heidelberg, 2013. Springer-Verlag. ISBN 978-3-642-36972-8. doi: 10.1007/978-3-642-36973-5_20. URL http://dx.doi.org/10.1007/978-3-642-36973-5_20. 112

ilad Shokouhi and Luo Si. Federated search. In *FNTIR*, 2011. 144

Bernard Jansen, Major Bernard, J. Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing & Management*, 36:207–227, March 2000. 23

Bernard J Jansen. Search log analysis: What it is, what's been done, how to do it. *Library & information science research*, 28(3):407–432, 2006. 20

Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. Determining the informational, navigational, and transactional intent of web queries. *Inf. Process. Manage.*, 44(3):1251–1266, 2008. ISSN 0306-4573. doi: http://dx.doi.org/10.1016/j.ipm.2007.07.015. 27, 35

Eric C. Jensen, Steven M. Beitzel, Abdur Chowdhury, and Ophir Frieder. Query phrase suggestion from topically tagged session logs. In *FQAS*, pages 185–196, 2006. 122

Hanna Jochmann-Mannak, Theo Huibers, Leo Lentz, and Ted Sanders. Children searching information on the internet: Performance on children's interfaces compared to google. In Pavel Serdyukov, Djoerd Hiemstra, and Ian Ruthven, editors, *Towards Accessible Search Systems*. ACM, July 2010. URL http://doc.utwente.nl/72475/. 18

Hanna Jochmann-Mannak, Leo Lentz, Theo Huibers, and Ted Sanders. Three types of children's informational web sites: An inventory of design conventions. *Technical Communication*, 59(4):302–323, 2012. 18, 210

Rosie Jones and Kristina Lisa Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *CIKM '08*, pages 699–708, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. doi: http://doi.acm.org/10.1145/1458082.1458176. URL `http://doi.acm.org/10.1145/1458082.1458176`. 19, 21

Rosie Jones, Ravi Kumar, Bo Pang, and Andrew Tomkins. "I know what you did last summer": query logs and user privacy. In *CIKM '07*, pages 909–914, New York, NY, USA, 2007. ACM. 64

Robert Kail. *Children and Their Development*. Pearson Education. Pearson, USA, 2009. 53

Yvonne Kammerer and Maja Bohnacker. Children's web search with google: the effectiveness of natural language queries. In *Proceedings of the 11th International Conference on Interaction Design and Children*, IDC '12, pages 184–187, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1007-9. doi: 10.1145/2307096.2307121. URL `http://doi.acm.org/10.1145/2307096.2307121`. 17

Christian Kohlschütter, Paul-Alexandru Chirita, and Wolfgang Nejdl. Utility analysis for topically biased pagerank. In *WWW '07*, pages 1211–1212, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7. doi: 10.1145/1242572.1242770. URL `http://doi.acm.org/10.1145/1242572.1242770`. 111

Arlind Kopliku. Aggregated search: From information nuggets to aggregated documents. In *CORIA*, pages 495–502, 2009. 7

C. C. Kuhlthau. Inside the search process: Information seeking from the user's perspective, 1991a. 13

Carol C Kuhlthau. Inside the search process: Information seeking from the user's perspective. *JASIS*, 42(5):361–371, 1991b. 76

Carol Collier Kuhlthau. Accommodating the user's information search process: challenges for information retrieval system designers. *Bulletin of the American Society for Information Science and Technology*, 25(3):12–16, 1999. 76

Ravi Kumar and Andrew Tomkins. A characterization of online browsing behavior. In *WWW '10*, pages 561–570, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772748. URL `http://doi.acm.org/10.1145/1772690.1772748`. 2, 87, 89, 90, 92, 94, 97

Minxian Li, Jinhui Tang, Haojie Li, and Chunxia Zhao. Tag ranking by propagating relevance over tag and image graphs. In *Proceedings of the 4th International Conference on Internet Multimedia Computing and Service*, ICIMCS '12, pages 153–156, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1600-2. doi: 10.1145/2382336.2382380. URL `http://doi.acm.org/10.1145/2382336.2382380`. 111

Xirong Li, Cees G.M. Snoek, and Marcel Worring. Learning tag relevance by neighbor voting for social image retrieval. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, MIR '08, pages 180–187, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-312-9. doi: 10.1145/1460096.1460126. URL `http://doi.acm.org/10.1145/1460096.1460126`. 111

Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong-Jiang Zhang. Tag ranking. In *18th International World Wide Web Conference*, pages 351–351, April 2009. URL `http://www2009.eprints.org/36/`. 111, 113

S Livingstone, L. Haddon, A. Görzig, and K. Ólafsson. EU Kids Online Final Report II. The London School of Economics and Political Science, 2011. `http://www.children-go-online.net`. 1

Tom Zeller Jr M. Barbaro. A face is exposed for aol searcher no. 4417749. 20

Yunlong Ma, Hongfei Lin, and Yuan Lin. Selecting related terms in query-logs using two-stage simrank. In *CIKM '11*, pages 1969–1972, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0717-8. doi: http://doi.acm.org/10.1145/2063576.2063867. URL `http://doi.acm.org/10.1145/2063576.2063867`. 112

M Madden, A. Lenhart, M. Duggan, S Cortesi, and U. Gasser. Teens and Technology 2013. PEW Research Center, The Berkman Center for Internet and Society at Harvard University, 2013. `http://www.pewinternet.org`. 1

Qiaozhu Mei, Dengyong Zhou, and Kenneth Church. Query suggestion using hitting time. In *CIKM '08*, pages 469–478, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458145. URL `http://doi.acm.org/10.1145/1458082.1458145`. 109, 110

Roger K. Moore and Louis Ten Bosch. Modelling vocabulary growth from birth to young adulthood. In *INTERSPEECH*, pages 1727–1730, USA, 2009. Curran Associates, Inc. 79

Vanessa Murdock and Mounia Lalmas. Workshop on aggregated search. *SIGIR Forum*, 42 (2):80–83, 2008. ISSN 0163-5840. doi: http://doi.acm.org/10.1145/1480506.1480520. 7, 8, 44

Diane Nahl and Violet H Harada. Composing boolean search statements: Self-confidence, concept analysis, search logic, and errors. *School Library Media Quarterly*, 24:199–207, 1996. 1, 16, 57

Delia Neuman. High school students' use of databases: Results of a national delphi study. *Journal of the American Society for Information Science*, 46(4):284–298, 1995. 16

Ofcom. Uk children's media literacy: Research document, March 2010. URL `http://www.ofcom.org.uk/advice/media_literacy/medlitpub/medlitpubrss/ukchildrensml/ukchildrensml1.pdf`. 1, 27, 36, 43, 198

Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. In *InfoScale '06: Proceedings of the 1st international conference on Scalable information systems*, New York, NY, USA, 2006. ACM Press. ISBN 1595934286. doi: 10.1145/1146847.1146848. URL `http://portal.acm.org/citation.cfm?id=1146847.1146848`. 13, 19, 37, 149

Anette Pettersson, Ulf Olsson, and Christina Fjellström. Family life in grocery stores–a study of interaction between adults and children. *International Journal of Consumer Studies*, 28(4):317–328, 2004. 2

Nina Phan, Peter Bailey, and Ross Wilkinson. Understanding the relationship of information need specificity to search query length. In *SIGIR '07*, pages 709–710, New York, NY, USA, 2007. ACM. 23

Feng Qiu and Junghoo Cho. Automatic identification of user interest for personalized search. In *WWW '06*. 111

Sujith Ravi, Andrei Broder, Evgeniy Gabrilovich, Vanja Josifovski, Sandeep Pandey, and Bo Pang. Automatic generation of bid phrases for online advertising. In *WSDM '10*, pages 341–350, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-889-6. doi: 10.1145/1718487.1718530. URL `http://doi.acm.org/10.1145/1718487.1718530`. 112

Matt Richtel. Children fail to recognize online ads, study says, April 2011. `http://bits.blogs.nytimes.com/2011/04/21/children-fail-to-recognize-online-ads-study-says/?hpw`. 60

Matt Richtel and Miguel Helft. Facebook users who are under age raise concerns, March 2011. `http://www.nytimes.com/2011/03/12/technology/internet/12underage.html?pagewanted=1&_r=1`. 53

Soo Young Rieh and Hong Iris Xie. Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing & Management*, 42(3):751–768, 2006. 19, 20

Rajendra Kumar Roul and SK Sahay. An effective information retrieval for ambiguous query. *arXiv preprint arXiv:1204.1406*, 2012. 23

John Schacter, Gregory KWK Chung, and Aimée Dorr. Children's internet searching on complex problems: performance and process analyses. *Journal of the American society for Information Science*, 49(9):840–849, 1998. 58

Philipp Schaer. Better than their reputation? on the reliability of relevance assessments with students. In *CLEF*, pages 124–135, 2012. 159

Milad Shokouhi, Justin Zobel, Falk Scholer, and S. M. M. Tahaghoghi. Capturing collection size for distributed non-cooperative retrieval. In *SIGIR '06*, pages 316–323, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7. doi: 10.1145/1148170.1148227. URL `http://doi.acm.org/10.1145/1148170.1148227`. 160, 161

Luo Si and Jamie Callan. Relevant document distribution estimation method for resource selection. In *SIGIR '03*, pages 298–305, New York, NY, USA, 2003. ACM. ISBN 1-58113-646-3. doi: 10.1145/860435.860490. URL `http://doi.acm.org/10.1145/860435.860490`. 144, 147, 160, 161

Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999. ISSN 0163-5840. doi: http://doi.acm.org/10.1145/331403.331405. 19, 23, 35

Peter. Smith, Helen. Cowie, and Mark Blades. Understanding children's developments, 2011. 76

Paul Solomon. Children's information retrieval behavior: A case analysis of an opac. *Journal of the American Society for Information Science*, 44(5):245–264, 1993. 15

Amanda Spink, Dietmar Wolfram, Major B. J. Jansen, and Tefko Saracevic. Searching the web: the public and their queries. *J. Am. Soc. Inf. Sci. Technol.*, 52(3):226–234, 2001. ISSN 1532-2882. doi: http://dx.doi.org/10.1002/1097-4571(2000)9999:9999⟨::AID-ASI1591⟩3.3.CO;2-I. 19

Shanu Sushmita, Hideo Joho, and Mounia Lalmas. A task-based evaluation of an aggregated search interface. In *SPIRE*, pages 322–333, 2009. 148, 155

Shanu Sushmita, Hideo Joho, Mounia Lalmas, and Robert Villa. Factors affecting click-through behavior in aggregated search interfaces. In *CIKM '10*, pages 519–528, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0099-5. doi: 10.1145/1871437.1871506. URL `http://doi.acm.org/10.1145/1871437.1871506`. 148

Idan Szpektor, Aristides Gionis, and Yoelle Maarek. Improving recommendation for long-tail queries via templates. In *WWW '11*, pages 47–56, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0632-4. doi: http://doi.acm.org/10.1145/1963405.1963416. URL `http://doi.acm.org/10.1145/1963405.1963416`. 123, 134

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment in short strength detection informal text. *Journal of the American Society for Information Science and Technology*, 61:2544–2558, 2010. 76

Paul Thomas and David Hawking. Evaluation by comparing result sets in context. In *CIKM '06*, pages 94–101, New York, NY, USA, 2006. ACM. ISBN 1-59593-433-2. doi: 10.1145/1183614.1183632. URL http://doi.acm.org/10.1145/1183614.1183632. 148

Paul Thomas and David Hawking. Evaluating sampling methods for uncooperative collections. In *SIGIR '07*, pages 503–510, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277828. URL http://doi.acm.org/10.1145/1277741.1277828. 159

Claudio Tonzar, Lorella Lotto, and Remo Job. L2 vocabulary acquisition in children: Effects of learning method and cognate status. *Language Learning*, 59(3):623–646, 2009. URL http://doi.wiley.com/10.1111/j.1467-9922.2009.00519.x. 30

Tim Van de Cruys. Two multivariate generalizations of pointwise mutual information. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, DiSCo '11, pages 16–20, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 9781937284022. URL http://dl.acm.org/citation.cfm?id=2043121.2043124. 120, 121

Frans van der Sluis, Sergio Duarte Torres, Djoerd Hiemstra, Betsy van Dijk, and Frea Kruisinga. Visual exploration of health information for children. In *ECIR'11*, pages 788–792, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-20160-8. URL http://dl.acm.org/citation.cfm?id=1996889.1997004. 110

Chiara Vettori and Ornella Mich. Supporting deaf children's reading skills: the many challenges of text simplification. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, ASSETS '11, pages 283–284, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0920-2. doi: 10.1145/2049536.2049608. URL http://doi.acm.org/10.1145/2049536.2049608. 99

Luis von Ahn and Laura Dabbish. Designing games with a purpose. *Commun. ACM*, 51 (8):58–67, August 2008. ISSN 0001-0782. doi: 10.1145/1378704.1378719. URL http://doi.acm.org/10.1145/1378704.1378719. 146, 170, 171

V. A Walter. The information needs of children. *Advances in Librarianship*, pages 111–129, 1994. 13

Jian Wang and Brian D. Davison. Explorations in tag suggestion and query expansion. In *SSM '08: Proceeding of the 2008 ACM workshop on Search in social media*, pages 43–50, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-258-0. doi: http://doi.acm.org/10.1145/1458583.1458592. 107, 122

Xuanhui Wang and ChengXiang Zhai. Mining term association patterns from search logs for effective query reformulation. In *CIKM '08*, pages 479–488, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. doi: http://doi.acm.org/10.1145/1458082.1458147. 28

Eugene J Webb, Donald T Campbell, Richard D Schwartz, Lee Sechrest, and Janet Belew Grove. Nonreactive measures in the social sciences. 1981. 2

Ingmar Weber and Carlos Castillo. The demographics of web search. In *SIGIR '10*, pages 523–530, New York, NY, USA, 2010. ACM. 49, 51

Ingmar Weber and Alejandro Jaimes. Who uses web search for what: and how. In *WSDM '11*, pages 15–24, New York, NY, USA, 2011. ACM. 49, 51, 53, 64

Ingmar Weber, Antti Ukkonen, and Aristides Gionis. Answers, not links: extracting tips from yahoo! answers to address how-to web queries. In *WSDM '12*, pages 613–622, New York, NY, USA, 2012. ACM. 68

Robert Wetzker, Carsten Zimmermann, and Christian Bauckhage. Analyzing social bookmarking systems: A del.icio.us cookbook. In *Mining Social Data (MSoDa) Workshop Proceedings*, pages 26–30. ECAI 2008, July 2008. URL `http://robertwetzker.com/wp-content/uploads/2008/06/wetzker_delicious_ecai2008_final.pdf`. 113, 119, 164

Ryen W. White, Peter Bailey, and Liwei Chen. Predicting user interests from contextual information. In *SIGIR '09*, pages 363–370, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1572005. URL `http://doi.acm.org/10.1145/1571941.1572005`. 42

Baoning Wu and Kumar Chellapilla. Extracting link spam using biased random walks from spam seed sets. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, AIRWeb '07, pages 37–44, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-732-2. doi: 10.1145/1244408.1244416. URL `http://doi.acm.org/10.1145/1244408.1244416`. 111, 118, 119, 125

Jingfang Xu, Sheng Wu, and Xing Li. Estimating collection size with logistic regression. In *SIGIR '07*, pages 789–790, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277910. URL `http://doi.acm.org/10.1145/1277741.1277910`. 159

Yusuke Yanbe, Adam Jatowt, Satoshi Nakamura, and Katsumi Tanaka. Can social bookmarking enhance search in the web? In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 107–116, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-644-8. doi: http://doi.acm.org/10.1145/1255175.1255198. 108, 116

Wei Vivian Zhang, Xiaofei He, Benjamin Rey, and Rosie Jones. Query rewriting using active learning for sponsored search. In *SIGIR '07*, pages 853–854, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7. doi: http://doi.acm.org/10.1145/1277741.1277942. 44

Xianchao Zhang, Bo Han, and Wenxin Liang. Automatic seed set expansion for trust propagation based anti-spamming algorithms. In *Proceedings of the eleventh international*

workshop on Web information and data management, WIDM '09, pages 31–38, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-808-7. doi: 10.1145/1651587.1651596. URL http://doi.acm.org/10.1145/1651587.1651596. 111

Ke Zhou, Ronan Cummins, Mounia Lalmas, and Joemon Jose. Evaluating large-scale distributed vertical search. In *Proceedings of the 9th workshop on Large-scale and distributed informational retrieval*, LSDS-IR '11, pages 9–14, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0959-2. doi: 10.1145/2064730.2064735. URL http://doi.acm.org/10.1145/2064730.2064735. 144

Guangyu Zhu, Shuicheng Yan, and Yi Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *Proceedings of the international conference on Multimedia*, MM '10, pages 461–470, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-933-6. doi: 10.1145/1873951.1874028. URL http://doi.acm.org/10.1145/1873951.1874028. 111

Xiaofei Zhu, Jiafeng Guo, Xueqi Cheng, Pan Du, and Hua-Wei Shen. A unified framework for recommending diverse and relevant queries. In *WWW '11*, pages 37–46, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0632-4. doi: 10.1145/1963405.1963415. URL http://doi.acm.org/10.1145/1963405.1963415. 195

Jinfeng Zhuang and Steven C.H. Hoi. A two-view learning approach for image tag ranking. In *WSDM '11*, pages 625–634, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0493-1. doi: 10.1145/1935826.1935913. URL http://doi.acm.org/10.1145/1935826.1935913. 111

SIKS Dissertation Series

2014-02 Fiona Tulinayo (RUN), *Combining System Dynamics with a Domain Modeling Method.*
2014-01 Nicola Barile (UU), *Studies in Learning Monotone Models from Data.*
2013-43 Marc Bron (UVA), *Exploration and Contextualization through Interaction and Concepts.*
2013-42 Léon Planken (TUD), *Algorithms for Simple Temporal Reasoning.*
2013-41 Jochem Liem (UVA), *Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning.*
2013-40 Pim Nijssen (UM), *Monte-Carlo Tree Search for Multi-Player Games.*
2013-39 Joop de Jong (TUD), *A Method for Enterprise Ontology based Design of Enterprise Information Systems.*
2013-38 Eelco den Heijer (VU), *Autonomous Evolutionary Art.*
2013-37 Dirk Börner (OUN), *Ambient Learning Displays.*
2013-36 Than Lam Hoang (TUe), *Pattern Mining in Data Streams.*
2013-35 Abdallah El Ali (UvA), *Minimal Mobile Human Computer Interaction.*
2013-34 Kien Tjin-Kam-Jet (UT), *Distributed Deep Web Search.*
2013-33 Qi Gao (TUD), *User Modeling and Personalization in the Microblogging Sphere.*
2013-32 Kamakshi Rajagopal (OUN), *Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development.*
2013-31 Dinh Khoa Nguyen (UvT), *Blueprint Model and Language for Engineering Cloud Applications.*
2013-30 Joyce Nakatumba (TUE), *Resource-Aware Business Process Management: Analysis and Support.*
2013-29 Iwan de Kok (UT), *Listening Heads.*
2013-28 Frans van der Sluis (UT), *When Complexity becomes Interesting: An Inquiry into the Information eXperience.*
2013-27 Mohammad Huq (UT), *Inference-based Framework Managing Data Provenance.*
2013-26 Alireza Zarghami (UT), *Architectural Support for Dynamic Homecare Service Provisioning.*
2013-25 Agnieszka Anna Latoszek-Berendsen (UM), *Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System.*
2013-24 Haitham Bou Ammar (UM), *Automated Transfer in Reinforcement Learning.*
2013-23 Patricio de Alencar Silva(UvT), *Value Activity Monitoring.*
2013-22 Tom Claassen (RUN), *Causal Discovery and Logic.*
2013-21 Sander Wubben (UvT), *Text-to-text generation by monolingual machine translation.*
2013-20 Katja Hofmann (UvA), *Fast and Reliable Online Learning to Rank for Information Retrieval.*
2013-19 Renze Steenhuizen (TUD), *Coordinated Multi-Agent Planning and Scheduling.*
2013-18 Jeroen Janssens (UvT), *Outlier Selection and One-Class Classification.*
2013-17 Koen Kok (VU), *The PowerMatcher: Smart Coordination for the Smart Electricity Grid.*
2013-16 Eric Kok (UU), *Exploring the practical benefits of argumentation in multi-agent deliberation.*
2013-15 Daniel Hennes (UM), *Multiagent Learning - Dynamic Games and Applications.*
2013-14 Jafar Tanha (UVA), *Ensemble Approaches to Semi-Supervised Learning Learning.*
2013-13 Mohammad Safiri(UT), *Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly.*
2013-12 Marian Razavian(VU), *Knowledge-driven Migration to Services.*
2013-11 Evangelos Pournaras(TUD), *Multi-level Reconfigurable Self-organization in Overlay Services.*
2013-10 Jeewanie Jayasinghe Arachchige(UvT), *A Unified Modeling Framework for Service Design..*
2013-09 Fabio Gori (RUN), *Metagenomic Data Analysis: Computational Methods and Applications.*
2013-08 Robbert-Jan Merk(VU), *Making enemies: cognitive modeling for opponent agents in fighter pilot simulators.*
2013-07 Giel van Lankveld (UvT), *Quantifying Individual Player Differences.*
2013-06 Romulo Goncalves(CWI), *The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience.*
2013-05 Dulce Pumareja (UT), *Groupware*

Requirements Evolutions Patterns.
2013-04 Chetan Yadati(TUD), *Coordinating autonomous planning and scheduling.*
2013-03 Szymon Klarman (VU), *Reasoning with Contexts in Description Logics.*
2013-02 Erietta Liarou (CWI), *MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing.*
2013-01 Viorel Milea (EUR), *News Analytics for Financial Decision Support.*
2012-51 Jeroen de Jong (TUD), *Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching.*
2012-50 Steven van Kervel (TUD), *Ontologogy driven Enterprise Information Systems Engineering.*
2012-49 Michael Kaisers (UM), *Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions.*
2012-48 Jorn Bakker (TUE), *Handling Abrupt Changes in Evolving Time-series Data.*
2012-47 Manos Tsagkias (UVA), *Mining Social Media: Tracking Content and Predicting Behavior.*
2012-46 Simon Carter (UVA), *Exploration and Exploitation of Multilingual Data for Statistical Machine Translation.*
2012-45 Benedikt Kratz (UvT), *A Model and Language for Business-aware Transactions.*
2012-44 Anna Tordai (VU), *On Combining Alignment Techniques.*
2012-43 Withdrawn, .
2012-42 Dominique Verpoorten (OU), *Reflection Amplifiers in self-regulated Learning.*
2012-41 Sebastian Kelle (OU), *Game Design Patterns for Learning.*
2012-40 Agus Gunawan (UvT), *Information Access for SMEs in Indonesia.*
2012-39 Hassan Fatemi (UT), *Risk-aware design of value and coordination networks.*
2012-38 Selmar Smit (VU), *Parameter Tuning and Scientific Testing in Evolutionary Algorithms.*
2012-37 Agnes Nakakawa (RUN), *A Collaboration Process for Enterprise Architecture Creation.*
2012-36 Denis Ssebugwawo (RUN), *Analysis and Evaluation of Collaborative Modeling Processes.*
2012-35 Evert Haasdijk (VU), *Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics.*
2012-34 Pavol Jancura (RUN), *Evolutionary analysis in PPI networks and applications.*
2012-33 Rory Sie (OUN), *Coalitions in Cooperation Networks (COCOON).*
2012-32 Wietske Visser (TUD), *Qualitative multi-criteria preference representation and reasoning.*
2012-31 Emily Bagarukayo (RUN), *A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure.*
2012-30 Alina Pommeranz (TUD), *Designing Human-Centered Systems for Reflective Decision Making.*
2012-29 Almer Tigelaar (UT), *Peer-to-Peer Information Retrieval.*
2012-28 Nancy Pascall (UvT), *Engendering Technology Empowering Women.*
2012-27 Hayrettin Gurkok (UT), *Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games.*
2012-26 Emile de Maat (UVA), *Making Sense of Legal Text.*
2012-25 Silja Eckartz (UT), *Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application.*
2012-24 Laurens van der Werff (UT), *Evaluation of Noisy Transcripts for Spoken Document Retrieval.*
2012-23 Christian Muehl (UT), *Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction.*
2012-22 Thijs Vis (UvT), *Intelligence, politie en veiligheidsdienst: verenigbare grootheden?.*
2012-21 Roberto Cornacchia (TUD), *Querying Sparse Matrices for Information Retrieval.*
2012-20 Ali Bahramisharif (RUN), *Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing.*
2012-19 Helen Schonenberg (TUE), *What's Next? Operational Support for Business Process Execution.*
2012-18 Eltjo Poort (VU), *Improving Solution Architecting Practices.*
2012-17 Amal Elgammal (UvT), *Towards a Comprehensive Framework for Business Process Compliance.*
2012-16 Fiemke Both (VU), *Helping people by understanding them - Ambient Agents supporting task execution and depression treatment.*
2012-15 Natalie van der Wal (VU), *Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes..*
2012-14 Evgeny Knutov(TUE), *Generic Adaptation Framework for Unifying Adaptive Web-based Systems.*
2012-13 Suleman Shahid (UvT), *Fun and Face: Exploring non-verbal expressions of emotion during playful interactions.*
2012-12 Kees van der Sluijs (TUE), *Model Driven Design and Data Integration in Semantic Web Information Systems.*
2012-11 J.C.B. Rantham Prabhakara (TUE), *Process Mining in the Large: Preprocessing, Discovery, and Diagnostics.*
2012-10 David Smits (TUE), *Towards a Generic Distributed Adaptive Hypermedia Environment.*
2012-09 Ricardo Neisse (UT), *Trust and Privacy Management Support for Context-Aware Service Platforms.*
2012-08 Gerben de Vries (UVA), *Kernel Methods for Vessel Trajectories.*
2012-07 Rianne van Lambalgen (VU), *When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions.*
2012-06 Wolfgang Reinhardt (OU), *Awareness Support for Knowledge Workers in Research Networks.*
2012-05 Marijn Plomp (UU), *Maturing*

Interorganisational Information Systems.

2012-04 Jurriaan Souer (UU), *Development of Content Management System-based Web Applications.*

2012-03 Adam Vanya (VU), *Supporting Architecture Evolution by Mining Software Repositories.*

2012-02 Muhammad Umair(VU), *Adaptivity, emotion, and Rationality in Human and Ambient Agent Models.*

2012-01 Terry Kakeeto (UvT), *Relationship Marketing for SMEs in Uganda.*

2011-49 Andreea Niculescu (UT), *Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality.*

2011-48 Mark Ter Maat (UT), *Response Selection and Turn-taking for a Sensitive Artificial Listening Agent.*

2011-47 Azizi Bin Ab Aziz(VU), *Exploring Computational Models for Intelligent Support of Persons with Depression.*

2011-46 Beibei Hu (TUD), *Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work.*

2011-45 Herman Stehouwer (UvT), *Statistical Language Models for Alternative Sequence Selection.*

2011-44 Boris Reuderink (UT), *Robust Brain-Computer Interfaces.*

2011-43 Henk van der Schuur (UU), *Process Improvement through Software Operation Knowledge.*

2011-42 Michal Sindlar (UU), *Explaining Behavior through Mental State Attribution.*

2011-41 Luan Ibraimi (UT), *Cryptographically Enforced Distributed Data Access Control.*

2011-40 Viktor Clerc (VU), *Architectural Knowledge Management in Global Software Development.*

2011-39 Joost Westra (UU), *Organizing Adaptation using Agents in Serious Games.*

2011-38 Nyree Lemmens (UM), *Bee-inspired Distributed Optimization.*

2011-37 Adriana Burlutiu (RUN), *Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference.*

2011-36 Erik van der Spek (UU), *Experiments in serious game design: a cognitive approach.*

2011-35 Maaike Harbers (UU), *Explaining Agent Behavior in Virtual Training.*

2011-34 Paolo Turrini (UU), *Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations.*

2011-33 Tom van der Weide (UU), *Arguing to Motivate Decisions.*

2011-32 Nees-Jan van Eck (EUR), *Methodological Advances in Bibliometric Mapping of Science.*

2011-31 Ludo Waltman (EUR), *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality.*

2011-30 Egon van den Broek (UT), *Affective Signal Processing (ASP): Unraveling the mystery of emotions.*

2011-29 Faisal Kamiran (TUE), *Discrimination-aware Classification.*

2011-28 Rianne Kaptein(UVA), *Effective Focused Retrieval by Exploiting Query Context and Document Structure.*

2011-27 Aniel Bhulai (VU), *Dynamic website optimization through autonomous management of design patterns.*

2011-26 Matthijs Aart Pontier (VU), *Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots.*

2011-25 Syed Waqar ul Qounain Jaffry (VU)), *Analysis and Validation of Models for Trust Dynamics.*

2011-24 Herwin van Welbergen (UT), *Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior.*

2011-23 Wouter Weerkamp (UVA), *Finding People and their Utterances in Social Media.*

2011-22 Junte Zhang (UVA), *System Evaluation of Archival Description and Access.*

2011-21 Linda Terlouw (TUD), *Modularization and Specification of Service-Oriented Systems.*

2011-20 Qing Gu (VU), *Guiding service-oriented software engineering - A view-based approach.*

2011-19 Ellen Rusman (OU), *The Mind ' s Eye on Personal Profiles.*

2011-18 Mark Ponsen (UM), *Strategic Decision-Making in complex games.*

2011-17 Jiyin He (UVA), *Exploring Topic Structure: Coherence, Diversity and Relatedness.*

2011-16 Maarten Schadd (UM), *Selective Search in Games of Different Complexity.*

2011-15 Marijn Koolen (UvA), *The Meaning of Structure: the Value of Link Evidence for Information Retrieval.*

2011-14 Milan Lovric (EUR), *Behavioral Finance and Agent-Based Artificial Markets.*

2011-13 Xiaoyu Mao (UvT), *Airport under Control. Multiagent Scheduling for Airport Ground Handling.*

2011-12 Carmen Bratosin (TUE), *Grid Architecture for Distributed Process Mining.*

2011-11 Dhaval Vyas (UT), *Designing for Awareness: An Experience-focused HCI Perspective.*

2011-10 Bart Bogaert (UvT), *Cloud Content Contention.*

2011-09 Tim de Jong (OU), *Contextualised Mobile Media for Learning.*

2011-08 Nieske Vergunst (UU), *BDI-based Generation of Robust Task-Oriented Dialogues.*

2011-07 Yujia Cao (UT), *Multimodal Information Presentation for High Load Human Computer Interaction.*

2011-06 Yiwen Wang (TUE), *Semantically-Enhanced Recommendations in Cultural Heritage.*

2011-05 Base van der Raadt (VU), *Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline..*

2011-04 Hado van Hasselt (UU), *Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference learning algorithms.*

2011-03 Jan Martijn van der Werf (TUE), *Compositional Design and Verification of Component-Based Information Systems.*

2011-02 Nick Tinnemeier(UU), *Organizing Agent Organizations. Syntax and Operational Semantics of an*

Organization-Oriented Programming Language.

2011-01 Botond Cseke (RUN), *Variational Algorithms for Bayesian Inference in Latent Gaussian Models.*

2010-53 Edgar Meij (UVA), *Combining Concepts and Language Models for Information Access.*

2010-52 Peter-Paul van Maanen (VU), *Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention.*

2010-51 Alia Khairia Amin (CWI), *Understanding and supporting information seeking tasks in multiple sources.*

2010-50 Bouke Huurnink (UVA), *Search in Audiovisual Broadcast Archives.*

2010-49 Jahn-Takeshi Saito (UM), *Solving difficult game positions.*

2010-48 Withdrawn, .

2010-47 Chen Li (UT), *Mining Process Model Variants: Challenges, Techniques, Examples.*

2010-46 Vincent Pijpers (VU), *e3alignment: Exploring Inter-Organizational Business-ICT Alignment.*

2010-45 Vasilios Andrikopoulos (UvT), *A theory and model for the evolution of software services.*

2010-44 Pieter Bellekens (TUE), *An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain.*

2010-43 Peter van Kranenburg (UU), *A Computational Approach to Content-Based Retrieval of Folk Song Melodies.*

2010-42 Sybren de Kinderen (VU), *Needs-driven service bundling in a multi-supplier setting - the computational e3-service approach.*

2010-41 Guillaume Chaslot (UM), *Monte-Carlo Tree Search.*

2010-40 Mark van Assem (VU), *Converting and Integrating Vocabularies for the Semantic Web.*

2010-39 Ghazanfar Farooq Siddiqui (VU), *Integrative modeling of emotions in virtual agents.*

2010-38 Dirk Fahland (TUE), *From Scenarios to components.*

2010-37 Niels Lohmann (TUE), *Correctness of services and their composition.*

2010-36 Jose Janssen (OU), *Paving the Way for Lifelong Learning; Facilitating competence development through a learning path specification.*

2010-35 Dolf Trieschnigg (UT), *Proof of Concept: Concept-based Biomedical Information Retrieval.*

2010-34 Teduh Dirgahayu (UT), *Interaction Design in Service Compositions.*

2010-33 Robin Aly (UT), *Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval.*

2010-32 Marcel Hiel (UvT), *An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems.*

2010-31 Victor de Boer (UVA), *Ontology Enrichment from Heterogeneous Sources on the Web.*

2010-30 Marieke van Erp (UvT), *Accessing Natural History - Discoveries in data cleaning, structuring, and retrieval.*

2010-29 Stratos Idreos(CWI), *Database Cracking: Towards Auto-tuning Database Kernels.*

2010-28 Arne Koopman (UU), *Characteristic Relational Patterns.*

2010-27 Marten Voulon (UL), *Automatisch contracteren.*

2010-26 Ying Zhang (CWI), *XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines.*

2010-25 Zulfiqar Ali Memon (VU), *Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective.*

2010-24 Dmytro Tykhonov, *Designing Generic and Efficient Negotiation Strategies.*

2010-23 Bas Steunebrink (UU), *The Logical Structure of Emotions.*

2010-22 Michiel Hildebrand (CWI), *End-user Support for Access to Heterogeneous Linked Data.*

2010-21 Harold van Heerde (UT), *Privacy-aware data management by means of data degradation.*

2010-20 Ivo Swartjes (UT), *Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative.*

2010-19 Henriette Cramer (UvA), *People's Responses to Autonomous and Adaptive Systems.*

2010-18 Charlotte Gerritsen (VU), *Caught in the Act: Investigating Crime by Agent-Based Simulation.*

2010-17 Spyros Kotoulas (VU), *Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications.*

2010-16 Sicco Verwer (TUD), *Efficient Identification of Timed Automata, theory and practice.*

2010-15 Lianne Bodenstaff (UT), *Managing Dependency Relations in Inter-Organizational Models.*

2010-14 Sander van Splunter (VU), *Automated Web Service Reconfiguration.*

2010-13 Gianluigi Folino (RUN), *High Performance Data Mining using Bio-inspired techniques.*

2010-12 Susan van den Braak (UU), *Sensemaking software for crime analysis.*

2010-11 Adriaan Ter Mors (TUD), *The world according to MARP: Multi-Agent Route Planning.*

2010-10 Rebecca Ong (UL), *Mobile Communication and Protection of Children.*

2010-09 Hugo Kielman (UL), *A Politiele gegevensverwerking en Privacy, Naar een effectieve waarborging.*

2010-08 Krzysztof Siewicz (UL), *Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments.*

2010-07 Wim Fikkert (UT), *Gesture interaction at a Distance.*

2010-06 Sander Bakkes (UvT), *Rapid Adaptation of Video Game AI.*

2010-05 Claudia Hauff (UT), *Predicting the Effectiveness of Queries and Retrieval Systems.*

2010-04 Olga Kulyk (UT), *Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments.*

2010-03 Joost Geurts (CWI), *A Document Engineering Model and Processing Framework for Multimedia documents.*

2010-02 Ingo Wassink (UT), *Work flows in Life Science.*
2010-01 Matthijs van Leeuwen (UU), *Patterns that Matter.*
2009-46 Loredana Afanasiev (UvA), *Querying XML: Benchmarks and Recursion.*
2009-45 Jilles Vreeken (UU), *Making Pattern Mining Useful.*
2009-44 Roberto Santana Tapia (UT), *Assessing Business-IT Alignment in Networked Organizations.*
2009-43 Virginia Nunes Leal Franqueira (UT), *Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients.*
2009-42 Toine Bogers (UvT), *Recommender Systems for Social Bookmarking.*
2009-41 Igor Berezhnyy (UvT), *Digital Analysis of Paintings.*
2009-40 Stephan Raaijmakers (UvT), *Multinomial Language Learning: Investigations into the Geometry of Language.*
2009-39 Christian Stahl (TUE, Humboldt-Universitaet zu Berlin), *Service Substitution – A Behavioral Approach Based on Petri Nets.*
2009-38 Riina Vuorikari (OU), *Tags and self-organisation: a metadata ecology for learning resources in a multilingual context.*
2009-37 Hendrik Drachsler (OUN), *Navigation Support for Learners in Informal Learning Networks.*
2009-36 Marco Kalz (OUN), *Placement Support for Learners in Learning Networks.*
2009-35 Wouter Koelewijn (UL), *Privacy en Politiegegevens; Over geautomatiseerde normatieve informatie-uitwisseling.*
2009-34 Inge van de Weerd (UU), *Advancing in Software Product Management: An Incremental Method Engineering Approach.*
2009-33 Khiet Truong (UT), *How Does Real Affect Affect Affect Recognition In Speech?.*
2009-32 Rik Farenhorst (VU) and Remco de Boer (VU), *Architectural Knowledge Management: Supporting Architects and Auditors.*
2009-31 Sofiya Katrenko (UVA), *A Closer Look at Learning Relations from Text.*
2009-30 Marcin Zukowski (CWI), *Balancing vectorized query execution with bandwidth-optimized storage.*
2009-29 Stanislav Pokraev (UT), *Model-Driven Semantic Integration of Service-Oriented Applications.*
2009-28 Sander Evers (UT), *Sensor Data Management with Probabilistic Models.*
2009-27 Christian Glahn (OU), *Contextual Support of social Engagement and Reflection on the Web.*
2009-26 Fernando Koch (UU), *An Agent-Based Model for the Development of Intelligent Mobile Services.*
2009-25 Alex van Ballegooij (CWI), *"RAM: Array Database Management through Relational Mapping".*
2009-24 Annerieke Heuvelink (VUA), *Cognitive Models for Training Simulations.*
2009-23 Peter Hofgesang (VU), *Modelling Web Usage in a Changing Environment.*
2009-22 Pavel Serdyukov (UT), *Search For Expertise: Going beyond direct evidence.*
2009-21 Stijn Vanderlooy (UM), *Ranking and Reliable Classification.*
2009-20 Bob van der Vecht (UU), *Adjustable Autonomy: Controling Influences on Decision Making.*
2009-19 Valentin Robu (CWI), *Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets.*
2009-18 Fabian Groffen (CWI), *Armada, An Evolving Database System.*
2009-17 Laurens van der Maaten (UvT), *Feature Extraction from Visual Data.*
2009-16 Fritz Reul (UvT), *New Architectures in Computer Chess.*
2009-15 Rinke Hoekstra (UVA), *Ontology Representation - Design Patterns and Ontologies that Make Sense.*
2009-14 Maksym Korotkiy (VU), *From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA).*
2009-13 Steven de Jong (UM), *Fairness in Multi-Agent Systems.*
2009-12 Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin), *Operating Guidelines for Services.*
2009-11 Alexander Boer (UVA), *Legal Theory, Sources of Law & the Semantic Web.*
2009-10 Jan Wielemaker (UVA), *Logic programming for knowledge-intensive interactive applications.*
2009-09 Benjamin Kanagwa (RUN), *Design, Discovery and Construction of Service-oriented Systems.*
2009-08 Volker Nannen (VU), *Evolutionary Agent-Based Policy Analysis in Dynamic Environments.*
2009-07 Ronald Poppe (UT), *Discriminative Vision-Based Recovery and Recognition of Human Motion.*
2009-06 Muhammad Subianto (UU), *Understanding Classification.*
2009-05 Sietse Overbeek (RUN), *Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality.*
2009-04 Josephine Nabukenya (RUN), *Improving the Quality of Organisational Policy Making using Collaboration Engineering.*
2009-03 Hans Stol (UvT), *A Framework for Evidence-based Policy Making Using IT.*
2009-02 Willem Robert van Hage (VU), *Evaluating Ontology-Alignment Techniques.*
2009-01 Rasa Jurgelenaite (RUN), *Symmetric Causal Independence Models.*
2008-35 Ben Torben Nielsen (UvT), *Dendritic morphologies: function shapes structure.*
2008-34 Jeroen de Knijf (UU), *Studies in Frequent Tree Mining.*
2008-33 Frank Terpstra (UVA), *Scientific Workflow Design; theoretical and practical issues.*
2008-32 Trung H. Bui (UT), *Toward Affective Dialogue Management using Partially Observable Markov Decision Processes.*
2008-31 Loes Braun (UM), *Pro-Active Medical Information Retrieval.*
2008-30 Wouter van Atteveldt (VU), *Semantic Network Analysis: Techniques for Extracting,*

Representing and Querying Media Content.
2008-29 Dennis Reidsma (UT), *Annotations and Subjective Machines - Of Annotators, Embodied Agents, Users, and Other Humans.*
2008-28 Ildiko Flesch (RUN), *On the Use of Independence Relations in Bayesian Networks.*
2008-27 Hubert Vogten (OU), *Design and Implementation Strategies for IMS Learning Design.*
2008-26 Marijn Huijbregts (UT), *Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled.*
2008-25 Geert Jonker (UU), *Efficient and Equitable Exchange in Air Traffic Management Plan Repair using Spender-signed Currency.*
2008-24 Zharko Aleksovski (VU), *Using background knowledge in ontology matching.*
2008-23 Stefan Visscher (UU), *Bayesian network models for the management of ventilator-associated pneumonia.*
2008-22 Henk Koning (UU), *Communication of IT-Architecture.*
2008-21 Krisztian Balog (UVA), *People Search in the Enterprise.*
2008-20 Rex Arendsen (UVA), *Geen bericht, goed bericht. Een onderzoek naar de effecten van de introductie van elektronisch berichtenverkeer met de overheid op de administratieve lasten van bedrijven..*
2008-19 Henning Rode (UT), *From Document to Entity Retrieval: Improving Precision and Performance of Focused Text Search.*
2008-18 Guido de Croon (UM), *Adaptive Active Vision.*
2008-17 Martin Op 't Land (TUD), *Applying Architecture and Ontology to the Splitting and Allying of Enterprises.*
2008-16 Henriette van Vugt (VU), *Embodied agents from a user's perspective.*
2008-15 Martijn van Otterlo (UT), *The Logic of Adaptive Behavior: Knowledge Representation and Algorithms for the Markov Decision Process Framework in First-Order Domains..*
2008-14 Arthur van Bunningen (UT), *Context-Aware Querying; Better Answers with Less Effort.*
2008-13 Caterina Carraciolo (UVA), *Topic Driven Access to Scientific Handbooks.*
2008-12 Jozsef Farkas (RUN), *A Semiotically Oriented Cognitive Model of Knowledge Representation.*
2008-11 Vera Kartseva (VU), *Designing Controls for Network Organizations: A Value-Based Approach.*
2008-10 Wauter Bosma (UT), *Discourse oriented summarization.*
2008-09 Christof van Nimwegen (UU), *The paradox of the guided user: assistance can be counter-effective.*
2008-08 Janneke Bolt (UU), *Bayesian Networks: Aspects of Approximate Inference.*
2008-07 Peter van Rosmalen (OU), *Supporting the tutor in the design and support of adaptive e-learning.*
2008-06 Arjen Hommersom (RUN), *On the Application of Formal Methods to Clinical Guidelines, an Artificial Intelligence Perspective.*
2008-05 Bela Mutschler (UT), *Modeling and simulating causal dependencies on process-aware information systems from a cost perspective.*
2008-04 Ander de Keijzer (UT), *Management of Uncertain Data - towards unattended integration.*
2008-03 Vera Hollink (UVA), *Optimizing hierarchical menus: a usage-based approach.*
2008-02 Alexei Sharpanskykh (VU), *On Computer-Aided Methods for Modeling and Analysis of Organizations.*
2008-01 Katalin Boer-Sorbán (EUR), *Agent-Based Simulation of Financial Markets: A modular, continuous-time approach.*
2007-25 Joost Schalken (VU), *Empirical Investigations in Software Process Improvement.*
2007-24 Georgina Ramírez Camps (CWI), *Structural Features in XML Retrieval.*
2007-23 Peter Barna (TUE), *Specification of Application Logic in Web Information Systems.*
2007-22 Zlatko Zlatev (UT), *Goal-oriented design of value and process models from patterns.*
2007-21 Karianne Vermaas (UU), *Fast diffusion and broadening use: A research on residential adoption and usage of broadband internet in the Netherlands between 2001 and 2005.*
2007-20 Slinger Jansen (UU), *Customer Configuration Updating in a Software Supply Network.*
2007-19 David Levy (UM), *Intimate relationships with artificial partners.*
2007-18 Bart Orriens (UvT), *On the development an management of adaptive business collaborations.*
2007-17 Theodore Charitos (UU), *Reasoning with Dynamic Networks in Practice.*
2007-16 Davide Grossi (UU), *Designing Invisible Handcuffs. Formal investigations in Institutions and Organizations for Multi-agent Systems.*
2007-15 Joyca Lacroix (UM), *NIM: a Situated Computational Memory Model.*
2007-14 Niek Bergboer (UM), *Context-Based Image Analysis.*
2007-13 Rutger Rienks (UT), *Meetings in Smart Environments; Implications of Progressing Technology.*
2007-12 Marcel van Gerven (RUN), *Bayesian Networks for Clinical Decision Support: A Rational Approach to Dynamic Decision-Making under Uncertainty.*
2007-11 Natalia Stash (TUE), *Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System.*
2007-10 Huib Aldewereld (UU), *Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols.*
2007-09 David Mobach (VU), *Agent-Based Mediated Service Negotiation.*
2007-08 Mark Hoogendoorn (VU), *Modeling of Change in Multi-Agent Organizations.*
2007-07 Natasa Jovanovic' (UT), *To Whom It May Concern - Addressee Identification in Face-to-Face Meetings.*
2007-06 Gilad Mishne (UVA), *Applied Text Analytics for Blogs.*
2007-05 Bart Schermer (UL), *Software Agents, Surveillance, and the Right to Privacy: a Legislative*

Framework for Agent-enabled Surveillance.
2007-04 Jurriaan van Diggelen (UU), *Achieving Semantic Interoperability in Multi-agent Systems: a dialogue-based approach.*
2007-03 Peter Mika (VU), *Social Networks and the Semantic Web.*
2007-02 Wouter Teepe (RUG), *Reconciling Information Exchange and Confidentiality: A Formal Approach.*
2007-01 Kees Leune (UvT), *Access Control and Service-Oriented Architectures.*
2006-28 Borkur Sigurbjornsson (UVA), *Focused Information Access using XML Element Retrieval.*
2006-27 Stefano Bocconi (CWI), *Vox Populi: generating video documentaries from semantically annotated media repositories.*
2006-26 Vojkan Mihajlovic (UT), *Score Region Algebra: A Flexible Framework for Structured Information Retrieval.*
2006-25 Madalina Drugan (UU), *Conditional log-likelihood MDL and Evolutionary MCMC.*
2006-24 Laura Hollink (VU), *Semantic Annotation for Retrieval of Visual Resources.*
2006-23 Ion Juvina (UU), *Development of Cognitive Model for Navigating on the Web.*
2006-22 Paul de Vrieze (RUN), *Fundaments of Adaptive Personalisation.*
2006-21 Bas van Gils (RUN), *Aptness on the Web.*
2006-20 Marina Velikova (UvT), *Monotone models for prediction in data mining.*
2006-19 Birna van Riemsdijk (UU), *Cognitive Agent Programming: A Semantic Approach.*
2006-18 Valentin Zhizhkun (UVA), *Graph transformation for Natural Language Processing.*
2006-17 Stacey Nagata (UU), *User Assistance for Multitasking with Interruptions on a Mobile Device.*
2006-16 Carsten Riggelsen (UU), *Approximation Methods for Efficient Learning of Bayesian Networks.*
2006-15 Rainer Malik (UU), *CONAN: Text Mining in the Biomedical Domain.*
2006-14 Johan Hoorn (VU), *Software Requirements: Update, Upgrade, Redesign - towards a Theory of Requirements Change.*
2006-13 Henk-Jan Lebbink (UU), *Dialogue and Decision Games for Information Exchanging Agents.*
2006-12 Bert Bongers (VU), *Interactivation - Towards an e-cology of people, our technological environment, and the arts.*
2006-11 Joeri van Ruth (UT), *Flattening Queries over Nested Data Types.*
2006-10 Ronny Siebes (VU), *Semantic Routing in Peer-to-Peer Systems.*
2006-09 Mohamed Wahdan (UM), *Automatic Formulation of the Auditor's Opinion.*
2006-08 Eelco Herder (UT), *Forward, Back and Home Again - Analyzing User Behavior on the Web.*
2006-07 Marko Smiljanic (UT), *XML schema matching – balancing efficiency and effectiveness by means of clustering.*
2006-06 Ziv Baida (VU), *Software-aided Service Bundling - Intelligent Methods & Tools for Graphical Service Modeling.*
2006-05 Cees Pierik (UU), *Validation Techniques for Object-Oriented Proof Outlines.*
2006-04 Marta Sabou (VU), *Building Web Service Ontologies.*
2006-03 Noor Christoph (UVA), *The role of metacognitive skills in learning to solve problems.*
2006-02 Cristina Chisalita (VU), *Contextual issues in the design and use of information technology in organizations.*
2006-01 Samuil Angelov (TUE), *Foundations of B2B Electronic Contracting.*
2005-21 Wijnand Derks (UT), *Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics.*
2005-20 Cristina Coteanu (UL), *Cyber Consumer Law, State of the Art and Perspectives.*
2005-19 Michel van Dartel (UM), *Situated Representation.*
2005-18 Danielle Sent (UU), *Test-selection strategies for probabilistic networks.*
2005-17 Boris Shishkov (TUD), *Software Specification Based on Re-usable Business Components.*
2005-16 Joris Graaumans (UU), *Usability of XML Query Languages.*
2005-15 Tibor Bosse (VU), *Analysis of the Dynamics of Cognitive Processes.*
2005-14 Borys Omelayenko (VU), *Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics.*
2005-13 Fred Hamburg (UL), *Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen.*
2005-12 Csaba Boer (EUR), *Distributed Simulation in Industry.*
2005-11 Elth Ogston (VU), *Agent Based Matchmaking and Clustering - A Decentralized Approach to Search.*
2005-10 Anders Bouwer (UVA), *Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments.*
2005-09 Jeen Broekstra (VU), *Storage, Querying and Inferencing for Semantic Web Languages.*
2005-08 Richard Vdovjak (TUE), *A Model-driven Approach for Building Distributed Ontology-based Web Applications.*
2005-07 Flavius Frasincar (TUE), *Hypermedia Presentation Generation for Semantic Web Information Systems.*
2005-06 Pieter Spronck (UM), *Adaptive Game AI.*
2005-05 Gabriel Infante-Lopez (UVA), *Two-Level Probabilistic Grammars for Natural Language Parsing.*
2005-04 Nirvana Meratnia (UT), *Towards Database Support for Moving Object data.*
2005-03 Franc Grootjen (RUN), *A Pragmatic Approach to the Conceptualisation of Language.*
2005-02 Erik van der Werf (UM)), *AI techniques for the game of Go.*
2005-01 Floor Verdenius (UVA), *Methodological Aspects of Designing Induction-Based Applications.*
2004-20 Madelon Evers (Nyenrode), *Learning from Design: facilitating multidisciplinary design teams.*
2004-19 Thijs Westerveld (UT), *Using generative probabilistic models for multimedia retrieval.*

2004-18 Vania Bessa Machado (UvA), *Supporting the Construction of Qualitative Knowledge Models.*
2004-17 Mark Winands (UM), *Informed Search in Complex Games.*
2004-16 Federico Divina (VU), *Hybrid Genetic Relational Search for Inductive Learning.*
2004-15 Arno Knobbe (UU), *Multi-Relational Data Mining.*
2004-14 Paul Harrenstein (UU), *Logic in Conflict. Logical Explorations in Strategic Equilibrium.*
2004-13 Wojciech Jamroga (UT), *Using Multiple Models of Reality: On Agents who Know how to Play.*
2004-12 The Duy Bui (UT), *Creating emotions and facial expressions for embodied agents.*
2004-11 Michel Klein (VU), *Change Management for Distributed Ontologies.*
2004-10 Suzanne Kabel (UVA), *Knowledge-rich indexing of learning-objects.*
2004-09 Martin Caminada (VU), *For the Sake of the Argument; explorations into argument-based reasoning.*
2004-08 Joop Verbeek(UM), *Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politile gegevensuitwisseling en digitale expertise.*
2004-07 Elise Boltjes (UM), *Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral voor meisjes.*
2004-06 Bart-Jan Hommes (TUD), *The Evaluation of Business Process Modeling Techniques.*
2004-05 Viara Popova (EUR), *Knowledge discovery and monotonicity.*
2004-04 Chris van Aart (UVA), *Organizational Principles for Multi-Agent Architectures.*
2004-03 Perry Groot (VU), *A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving.*
2004-02 Lai Xu (UvT), *Monitoring Multi-party Contracts for E-business.*
2004-01 Virginia Dignum (UU), *A Model for Organizational Interaction: Based on Agents, Founded in Logic.*
2003-18 Levente Kocsis (UM), *Learning Search Decisions.*
2003-17 David Jansen (UT), *Extensions of Statecharts with Probability, Time, and Stochastic Timing.*
2003-16 Menzo Windhouwer (CWI), *Feature Grammar Systems - Incremental Maintenance of Indexes to Digital Media Warehouses.*
2003-15 Mathijs de Weerdt (TUD), *Plan Merging in Multi-Agent Systems.*
2003-14 Stijn Hoppenbrouwers (KUN), *Freezing Language: Conceptualisation Processes across ICT-Supported Organisations.*
2003-13 Jeroen Donkers (UM), *Nosce Hostem - Searching with Opponent Models.*
2003-12 Roeland Ordelman (UT), *Dutch speech recognition in multimedia information retrieval.*
2003-11 Simon Keizer (UT), *Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks.*
2003-10 Andreas Lincke (UvT), *Electronic Business Negotiation: Some experimental studies on the interaction between medium, innovation context and culture.*
2003-09 Rens Kortmann (UM), *The resolution of visually guided behaviour.*
2003-08 Yongping Ran (UM), *Repair Based Scheduling.*
2003-07 Machiel Jansen (UvA), *Formal Explorations of Knowledge Intensive Tasks.*
2003-06 Boris van Schooten (UT), *Development and specification of virtual environments.*
2003-05 Jos Lehmann (UVA), *Causation in Artificial Intelligence and Law - A modelling approach.*
2003-04 Milan Petkovic (UT), *Content-Based Video Retrieval Supported by Database Technology.*
2003-03 Martijn Schuemie (TUD), *Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy.*
2003-02 Jan Broersen (VU), *Modal Action Logics for Reasoning About Reactive Systems.*
2003-01 Heiner Stuckenschmidt (VU), *Ontology-Based Information Sharing in Weakly Structured Environments.*
2002-17 Stefan Manegold (UVA), *Understanding, Modeling, and Improving Main-Memory Database Performance.*
2002-16 Pieter van Langen (VU), *The Anatomy of Design: Foundations, Models and Applications.*
2002-15 Rik Eshuis (UT), *Semantics and Verification of UML Activity Diagrams for Workflow Modelling.*
2002-14 Wieke de Vries (UU), *Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems.*
2002-13 Hongjing Wu (TUE), *A Reference Architecture for Adaptive Hypermedia Applications.*
2002-12 Albrecht Schmidt (Uva), *Processing XML in Database Systems.*
2002-11 Wouter C.A. Wijngaards (VU), *Agent Based Modelling of Dynamics: Biological and Organisational Applications.*
2002-10 Brian Sheppard (UM), *Towards Perfect Play of Scrabble.*
2002-09 Willem-Jan van den Heuvel(KUB), *Integrating Modern Business Applications with Objectified Legacy Systems.*
2002-08 Jaap Gordijn (VU), *Value Based Requirements Engineering: Exploring Innovative E-Commerce Ideas.*
2002-07 Peter Boncz (CWI), *Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications.*
2002-06 Laurens Mommers (UL), *Applied legal epistemology; Building a knowledge-based ontology of the legal domain.*
2002-05 Radu Serban (VU), *The Private Cyberspace Modeling Electronic Environments inhabited by Privacy-concerned Agents.*
2002-04 Juan Roberto Castelo Valdueza (UU), *The Discrete Acyclic Digraph Markov Model in Data Mining.*
2002-03 Henk Ernst Blok (UT), *Database Optimization Aspects for Information Retrieval.*
2002-02 Roelof van Zwol (UT), *Modelling and searching web-based document collections.*
2002-01 Nico Lassing (VU), *Architecture-Level*

Modifiability Analysis.

2001-9 Pieter Jan 't Hoen (RUL), *Towards Distributed Development of Large Object-Oriented Models, Views of Packages as Classes.*

2001-8 Pascal van Eck (VU), *A Compositional Semantic Structure for Multi-Agent Systems Dynamics..*

2001-7 Bastiaan Schonhage (VU), *Diva: Architectural Perspectives on Information Visualization.*

2001-6 Martijn van Welie (VU), *Task-based User Interface Design.*

2001-5 Jacco van Ossenbruggen (VU), *Processing Structured Hypermedia: A Matter of Style.*

2001-4 Evgueni Smirnov (UM), *Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets.*

2001-3 Maarten van Someren (UvA), *Learning as problem solving.*

2001-2 Koen Hindriks (UU), *Agent Programming Languages: Programming with Mental Models.*

2001-11 Tom M. van Engers (VUA), *Knowledge Management: The Role of Mental Models in Business Systems Design.*

2001-10 Maarten Sierhuis (UvA), *Modeling and Simulating Work Practice BRAHMS: a multiagent modeling and simulation language for work practice analysis and design.*

2001-1 Silja Renooij (UU), *Qualitative Approaches to Quantifying Probabilistic Networks.*

2000-9 Florian Waas (CWI), *Principles of Probabilistic Query Optimization.*

2000-8 Veerle Coup, (EUR), *Sensitivity Analyis of Decision-Theoretic Networks.*

2000-7 Niels Peek (UU), *Decision-theoretic Planning of Clinical Patient Management.*

2000-6 Rogier van Eijk (UU), *Programming Languages for Agent Communication.*

2000-5 Ruud van der Pol (UM), *Knowledge-based Query Formulation in Information Retrieval..*

2000-4 Geert de Haan (VU), *ETAG, A Formal Model of Competence Knowledge for User Interface Design.*

2000-3 Carolien M.T. Metselaar (UVA), *Sociaal-organisatorische gevolgen van kennistechnologie; een procesbenadering en actorperspectief..*

2000-2 Koen Holtman (TUE), *Prototyping of CMS Storage Management.*

2000-11 Jonas Karlsson (CWI), *Scalable Distributed Data Structures for Database Management.*

2000-10 Niels Nes (CWI), *Image Database Management System Design Considerations, Algorithms and Architecture.*

2000-1 Frank Niessink (VU), *Perspectives on Improving Software Maintenance.*

1999-8 Jacques H.J. Lenting (UM), *Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation..*

1999-7 David Spelt (UT), *Verification support for object database design.*

1999-6 Niek J.E. Wijngaards (VU), *Re-design of compositional systems.*

1999-5 Aldo de Moor (KUB), *Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems.*

1999-4 Jacques Penders (UM), *The practical Art of Moving Physical Objects.*

1999-3 Don Beal (UM), *The Nature of Minimax Search.*

1999-2 Rob Potharst (EUR), *Classification using decision trees and neural nets.*

1999-1 Mark Sloof (VU), *Physiology of Quality Change Modelling; Automated modelling of Quality Change of Agricultural Products.*

1998-5 E.W.Oskamp (RUL), *Computerondersteuning bij Straftoemeting.*

1998-4 Dennis Breuker (UM), *Memory versus Search in Games.*

1998-3 Ans Steuten (TUD), *A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective.*

1998-2 Floris Wiesman (UM), *Information Retrieval by Graphically Browsing Meta-Information.*

1998-1 Johan van den Akker (CWI), *DEGAS - An Active, Temporal Database of Autonomous Objects.*