Very similar items lost in the Web:

An investigation of deduplication by *Google Web Search* and other search engines

- Wouter.Mettrop@cwi.nl

CWI, Amsterdam, The Netherlands

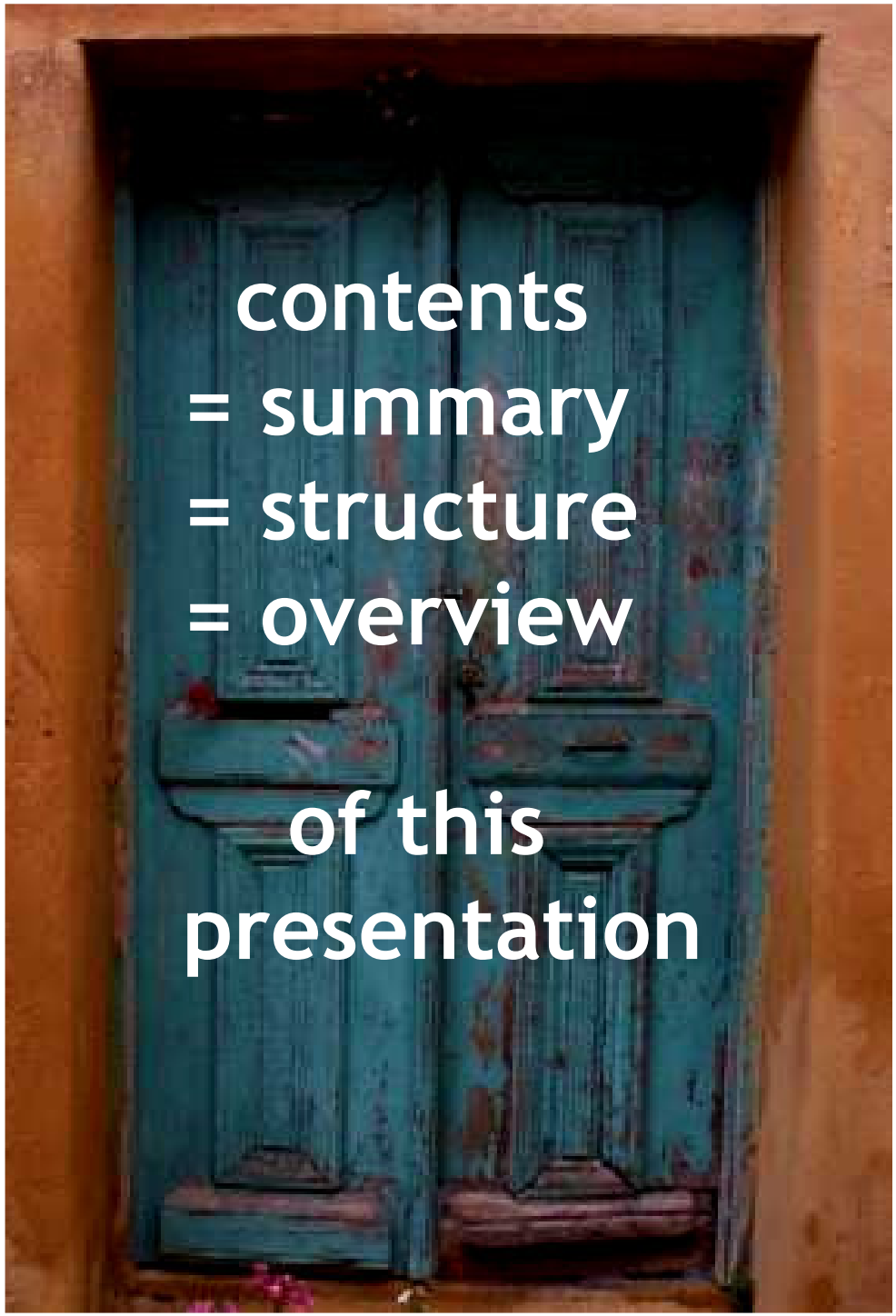
- Paul.Nieuwenhuysen@vub.ac.be

Vrije Universiteit Brussel, and *Universiteit Antwerpen*, Belgium

- Hanneke Smulders

Infomare Consultancy, The Netherlands

Prepared for a presentation at Information Online, in Sydney, 2007



Classical structure:

1. Introduction
2. Problem statements
3. Experimental procedure
4. Results
5. Discussion
6. Conclusion
   & recommendations

Introduction
Duplicate files

Many computer files that carry documents, images, multimedia, programs
are present in personal information systems, organizations, intranets, the Internet and the Web, …
in more than one copy or they are very similar to other files

Introduction
Duplicates on the Web exist

Investigation proved that about 30% of all Web pages are very similar to other pages of the remaining 70% and that about 20% are virtually identical to other pages on the Web.

Introduction:
Duplicates cause problems

1. **Storage of information:**
   - » Duplicate files consume memory and processing power of computers.
   - » This forms a challenge for information retrieval systems.
   - » Furthermore, as an increasing number of people create, copy, store and distribute files, this challenge gets more important.

Introduction:
Duplicates cause problems

2. Retrieval of information:

» What is worse:
Users loose time in locating the file that is the most appropriate or original or authentic or recent, wading through duplicates and near-duplicates.

Introduction:
Deduplication may be useful

- To help users in view of the many copies, duplicates, or very similar files, Web search engines can apply deduplication.
- ***Deduplication*** helps the user to identify duplicates by presenting only 1 or a few representatives instead of the whole cluster.
- So the user can review the results faster.

Purpose of this investigation

The investigation reported here was motivated by the central problem:

In which ways is the user confronted with the various ways in which Web search engines handle very similar documents?

Problem statement 1

1. Do the important, popular Web search engines offer their users results that have been deduplicated in some way?

Problem statement 2

2. How do similar documents show up in search results?

   Is this presentation constant over time?
   How often do changes occur?

Problem statement 3

3. Is the user confronted with deduplication by various Web search engines in the same way?

Problem statement 4

4. How stable and predictable is the deduplication function of Web search engines?

Problem statement 5

5. How should a user take deduplication into account?

Justification - these are relevant research questions because …

- a large part of files on the WWW are very similar
- clarification is desired for expert users of search engines in quantitative studies (informetrics)
- clarification is desired for any information searcher, as documents that are similar for computers can carry different meanings for a human reader
- WWW search engines have become quite important information systems with a huge user community

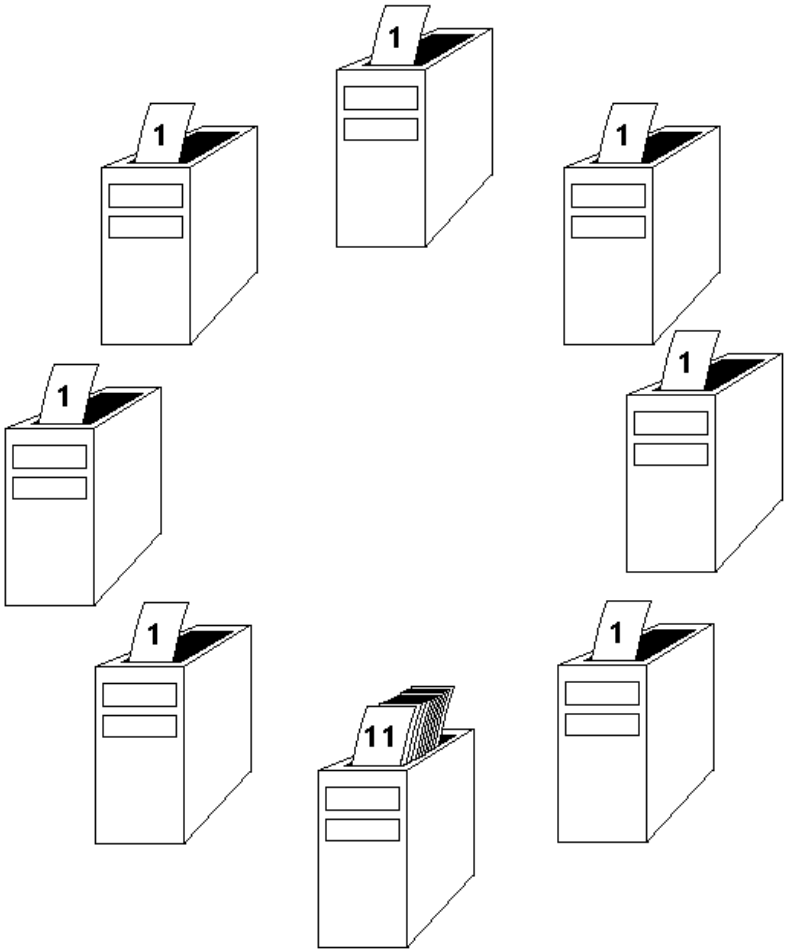
Experimental procedure

- We have performed experiments with very similar test documents.

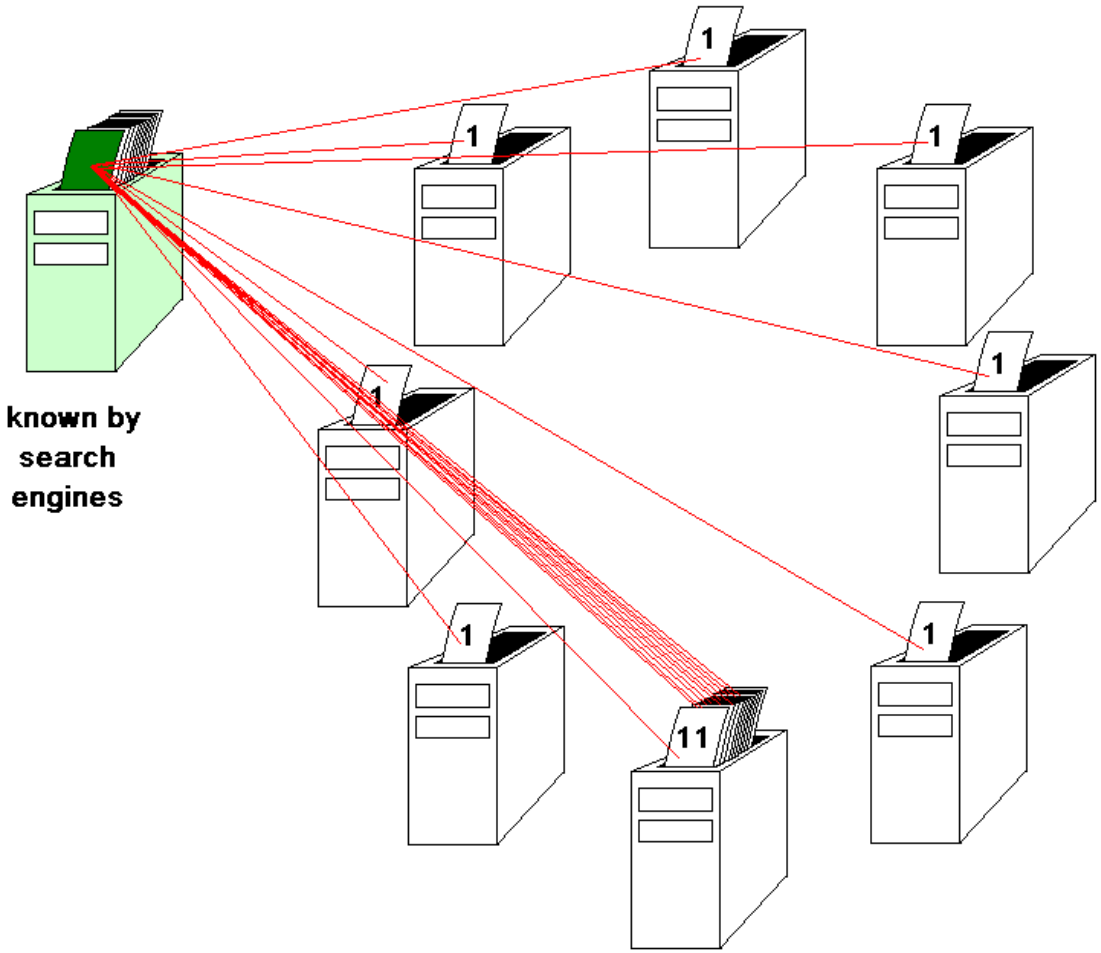
- We constructed a test document and 17 variations of this document.
  Differences among our test documents were made in
  the HTML title,
  body text and
  filename.

- We used 8 different WWW servers in 2 countries.

Experimental procedure



Experimental procedure



Experimental procedure
Web search engines investigated

- *Alltheweb*
- *AltaVista*
- *Ask*
- *Google Web Search*
- *Lycos*
- *MSN*
- *Teoma*
- *Yahoo!*

Experimental procedure

- The test documents were searched with one specific *content query* repeated every hour during September - October 2005.

- Every investigated Web search engine has been queried 430 times with the content query.

Experimental procedure

- Also they were queried simultaneously with 18 what we call "URL queries".
  These are queries that search for (a part of) the URL of the 18 test documents, using the possibilities that the particular search engine offers.
- We name the test documents retrieved by all 19 simultaneously submitted queries
  “known test documents”.
- The total number of queries submitted is 28886.

Results (1)

ASK TEOMA LYCOS

No deduplication

Results (2)

GOOGLE WEB SEARCH YAHOO!

Partial deduplication, user can ask for hidden documents

ASK TEOMA LYCOS

No deduplication

Information ONLINE 2007

Results (3)

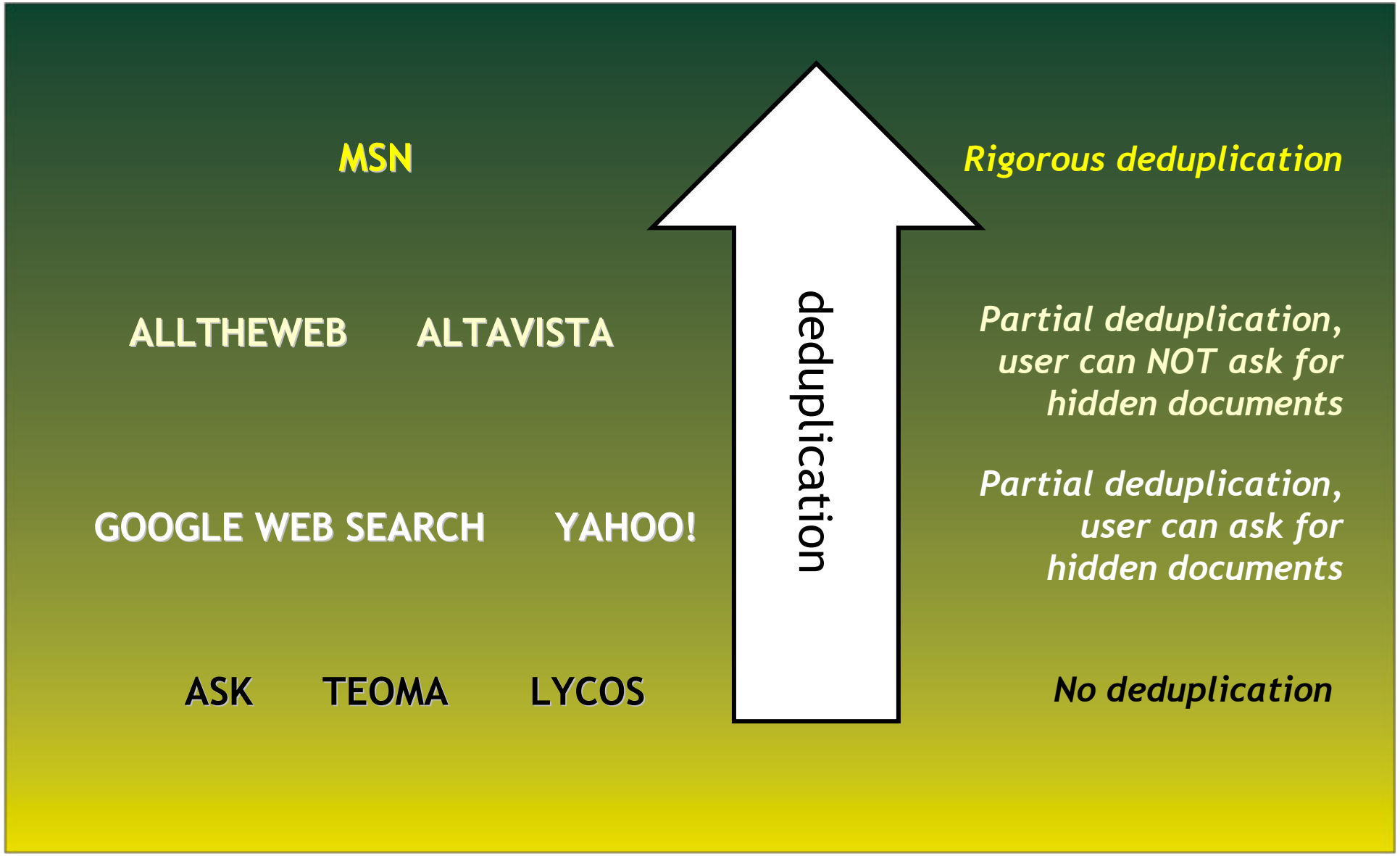
ALLTHEWEB ALTAVISTA	*Partial deduplication, user can NOT ask for hidden documents*
GOOGLE WEB SEARCH YAHOO!	*Partial deduplication, user can ask for hidden documents*
ASK TEOMA LYCOS	*No deduplication*

Information ONLINE 2007

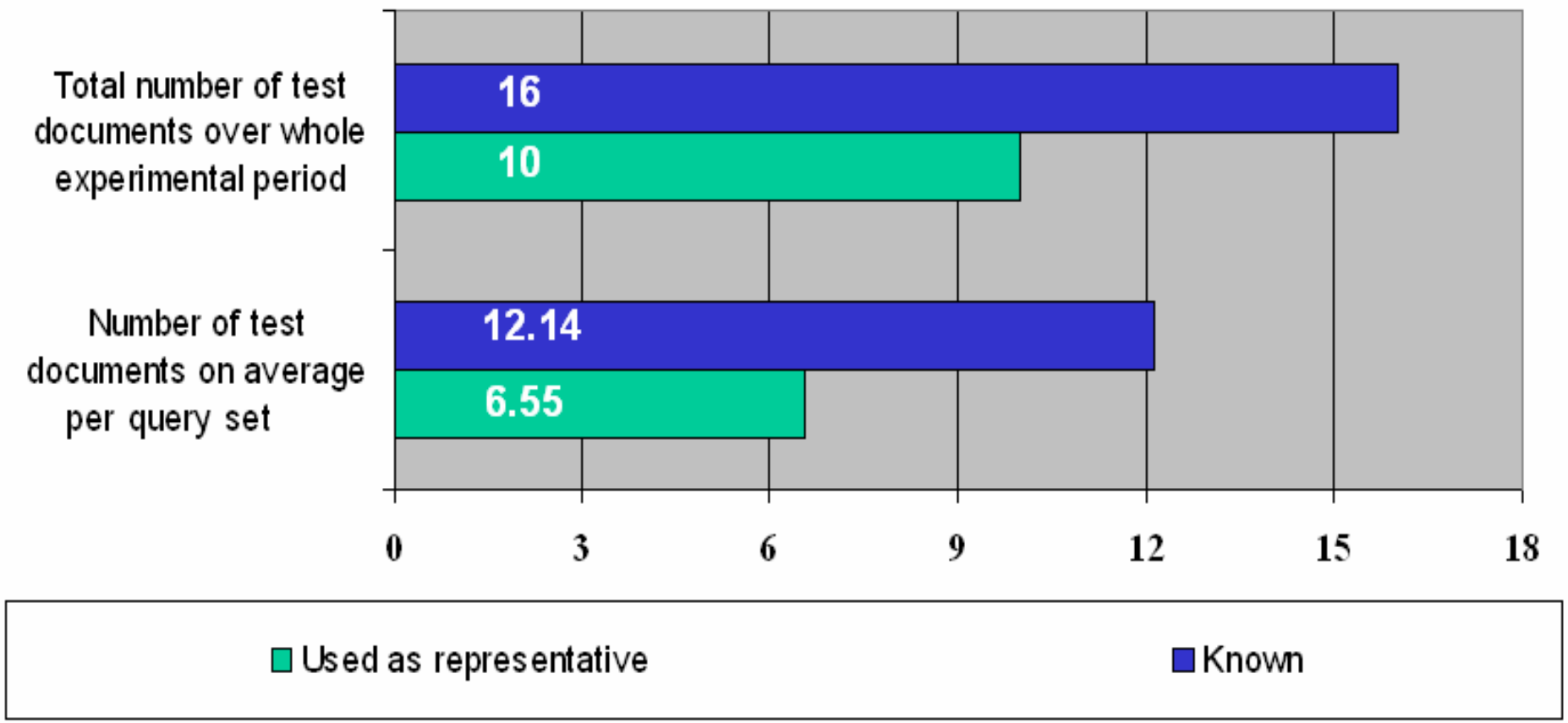
Results (4)

MSN	*Rigorous deduplication*
ALLTHEWEB ALTAVISTA	*Partial deduplication, user can NOT ask for hidden documents*
GOOGLE WEB SEARCH YAHOO!	*Partial deduplication, user can ask for hidden documents*
ASK TEOMA LYCOS	*No deduplication*

Results (5)


MSN
Rigorous deduplication
ALLTHEWEB
ALTAVISTA
Partial deduplication, user can NOT ask for hidden documents
GOOGLE WEB SEARCH
YAHOO!
Partial deduplication, user can ask for hidden documents
ASK
TEOMA
LYCOS
No deduplication
deduplication

Results example: *numbers of test documents involved for Yahoo!*

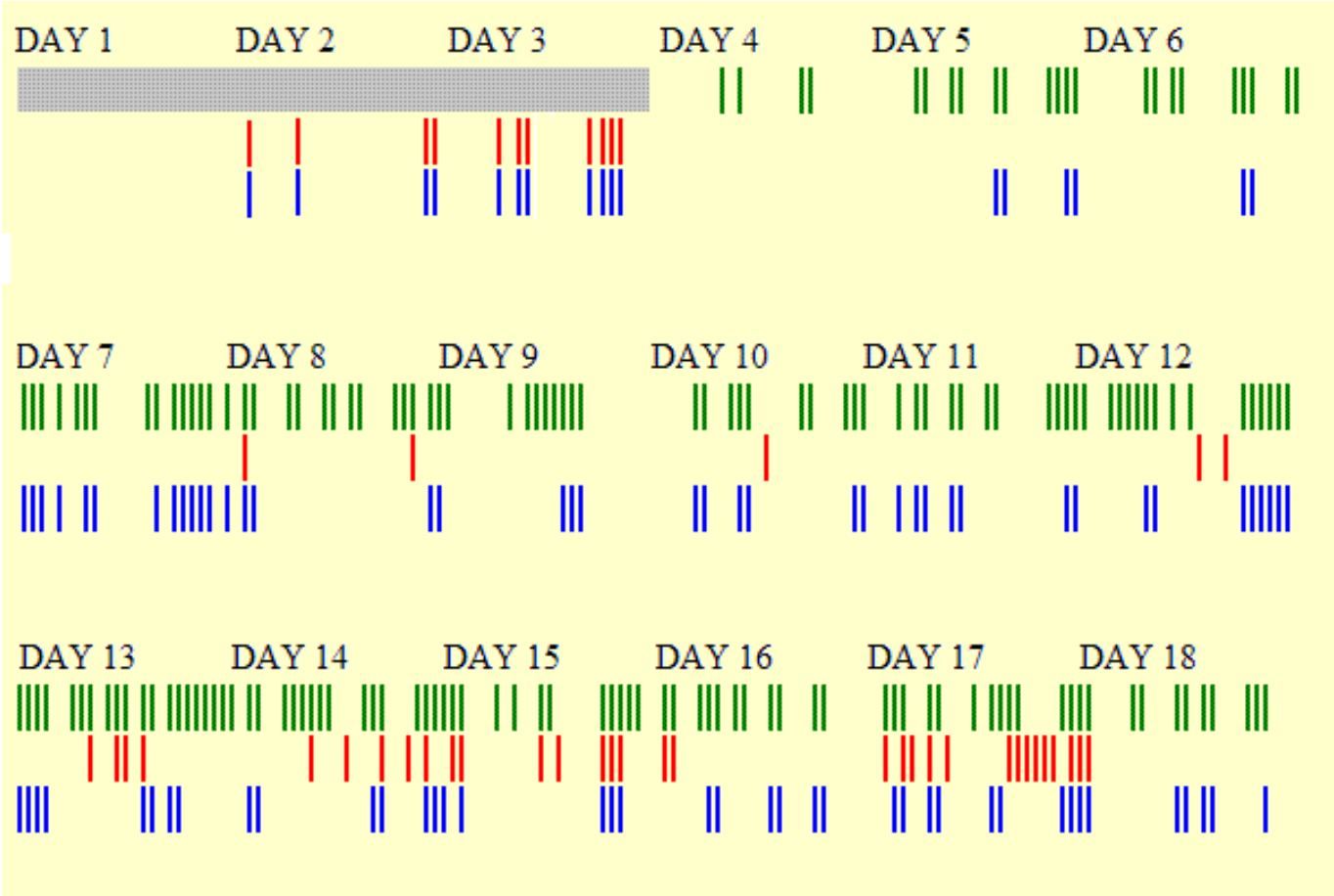


Results (6)
fluctuations over time

Fluctuations over time occurred in the result sets of some search engines,
i.e. queries do not always show the same set of test documents that was retrieved with the previous submission.

Results example: *fluctuations for Yahoo!*

URL
content
known



176
48
98

Results (7) - Quantitative presentation

	known documents, hidden per query set on average	deduplicated result sets with document fluctuations
MSN	84%	0%
Alltheweb	38%	5%
AltaVista	30%	5%
GoogleWebSearch	51%	0.1%
Yahoo!	46%	11%
AskJeeves	0%	0%
Lycos	0%	0%
Teoma	0%	0%

Results (7) - Quantitative presentation

	known documents, hidden per query set on average	deduplicated result sets with document fluctuations
MSN	84%	0%
Alltheweb	38%	5%
AltaVista	30%	5%
GoogleWebSearch	51%	0.1%
Yahoo!	46%	11%
AskJeeves	*0%*	*0%*
Lycos	*0%*	*0%*
Teoma	*0%*	*0%*

Results (7) - Quantitative presentation

	known documents, hidden per query set on average	deduplicated result sets with document fluctuations
MSN	84%	0%
Alltheweb	38%	5%
AltaVista	30%	5%
GoogleWebSearch	*51%*	*0.1%*
Yahoo!	*46%*	*11%*
AskJeeves	0%	0%
Lycos	0%	0%
Teoma	0%	0%

Results (7) - Quantitative presentation

	known documents, hidden per query set on average	deduplicated result sets with document fluctuations
MSN	84%	0%
Alltheweb	*38%*	*5%*
AltaVista	*30%*	*5%*
GoogleWebSearch	51%	0.1%
Yahoo!	46%	11%
AskJeeves	0%	0%
Lycos	0%	0%
Teoma	0%	0%

Results (7) - Quantitative presentation

	known documents, hidden per query set on average	deduplicated result sets with document fluctuations
MSN	*84%*	*0%*
Alltheweb	38%	5%
AltaVista	30%	5%
GoogleWebSearch	51%	0.1%
Yahoo!	46%	11%
AskJeeves	0%	0%
Lycos	0%	0%
Teoma	0%	0%

Screen shot of *Google* which has "omitted some entries"

Google

Web Images Groups News Froogle more »

"in deze uitgave van het spelregelboek zijn alle" Search Advanced Search Preferences

Web Results **1** - **1** of about **3** for **"in deze uitgave van het spelregelboek zijn alle"**. (**0.46** seconds)

Tip: Try removing quotes from your search to get more results.

Nieuwe pagina 0
... de overtreding plaatsvindt. **In deze uitgave van het spelregelboek zijn alle** wijzigingen tot en met 2001 verwerkt. Degenen die zich ...
www.rkwerp.nl/Spelregels.htm - 29k - Cached - Similar pages

In order to show you the most relevant results, we have omitted some entries very similar to the 1 already displayed.
If you like, you can repeat the search with the omitted results included.

Screen shot of *Yahoo!* which has "omitted some entries"

Yahoo! My Yahoo! Mail Welcome, **Guest** [Sign In] Search Home Help

YAHOO! SEARCH

Web | Images | Video | Audio | Directory | Local | News | Shopping | More »

"in deze uitgave van het spelregelboek zijn alle" Search

Answers My Web Search Services | Advanced Search Preferences

Search Results 1 - 3 of about 24 for "**in deze uitgave van het spelregelboek zijn alle**" - 0.77 sec. (About this page)

1. out (PDF)
   … **In deze uitgave van het spelregelboek zijn alle**. wijzigingen tot en met 2001 verwerkt …
   mikekoeleman.tripod.com/Spelregels_veldvoetbal.pdf - 184k - View as html - More from this site - Save

2. INLEIDING (PDF)
   … maar waar de overtreding plaatsvindt. **In deze uitgave van het spelregelboek zijn alle**. wijzigingen tot en met 2001 …
   www.vub.ac.be/BIBLIO/spelregels/spelregels_veldvoetbal.pdf - 214k - View as html - More from this site - Save

3. Regel a
   De spelregelwijzigingen die de FIFA in 2001 heeft doorgevoerd, zijn voor de werkgroep spelregels veldvoetbal aanleiding om een nieuw spelregelboek uit te geven. … **In deze uitgave van het spelregelboek zijn alle** wijzigingen tot en met 2001 verwerkt …
   diagonaal.info/regel-a.html - 2k - Cached - More from this site - Save

In order to show you the most relevant results, we have omitted some entries very similar to the ones already displayed.
If you like, you can repeat the search with the omitted results included.

Discussion:
The importance of our findings

- Real, authentic documents on their original server computer have to compete with “very similar” versions, which are made available by others on other servers.

- In reality documents are not abstract items:
  they can be concrete, real laws, regulations, price lists, scientific reports, political programs…
  so that NOT finding the more authentic document can have real consequences.

Discussion:
The importance of our findings

- Deduplication may complicate scientometric / bibliometric studies, quantitative studies of numbers of documents retrieved.

Discussion:
The importance of our findings

- Documents on their original server can be pushed away from the search results, by very similar competing documents on 1 or several other servers.

Discussion:
The importance of our findings

- Furthermore, documents that are "very similar" for a computer system can carry a substantially different meaning for a human user.
  A small change in a document may have large consequences for the meaning!

Conclusion
Recommendations

- Very similar documents are handled in different ways by different search engines.
- Deduplication takes place by several engines.
- Not only strict duplicates, but also very similar documents are omitted from search results.

- Enjoy this: when you don't want very similar documents in your search results.
  Then use a search engine that deduplicates rigorously

Conclusion
Recommendations

- But take deduplication into account when it is important to find
  - » the oldest, authentic, master version of a document;
  - » the newest, most recent version of a document;
  - » versions of a document with comments, corrections…
  - » in general: variations of documents

- In that case use a search engine that does not deduplicate or that allows you to view the omitted search results.

Conclusion
Recommendations

- Search engines that deduplicate partially show fluctuations in the search results over time.
- Searchers for a known item should be aware of this.