Implicit relevance feedback from a multi-step search process: a use of query-logs

Corrado Boscarino, Arjen P. de Vries, Vera Hollink, and Jacco van Ossenbruggen

Centrum Wiskunde & Informatica (CWI), Science Park 123,
1098 XG Amsterdam, The Netherlands
corrado@cwi.nl, arjen@acm.org, v.hollink@cwi.nl, jvosse@cwi.nl

Abstract. We evaluate the use of clickthrough information as implicit relevance feedback in sessions. We employ records of user interactions with a commercial news picture portal: issued queries, clicked images, and purchased content. Our study investigates how much of a session's search history (if any) should be used in a feedback loop. We assess the benefit of using clicked data as positive tokens of relevance to the task of estimating the probability of an image to be purchased. We find that a short history of past queries helps improve ranking, and that terms derived from clicked documents lead to a much higher effectiveness, while blind relevance feedback is not beneficial for the task.

1 Evidence of user interaction: Query Logs (QL)

Logs of queries issued and the subsequent interactions with the query results, briefly referred to as 'query logs' (QLs) in this paper, provide a basis to adapt a relevance model to reflect what we have learned about the user's information need. A set of QLs recorded when subscribers to Belga Picture[1] were searching for images to be purchased online, allows us 1) to investigate how valuable clicks are as source of (implicit) relevance feedback in a multi-step search session and 2) to observe how much search history (if any) may lead to an improvement in the ranking of what we believe to be a determinately relevant document: the picture that a user is known to have purchased at the end of a search session.

A QL registers, for each session, three types of user interactions: query submissions (Q), a possibly empty set of clicks (C) on the retrieved results, and, purchases (P); an anonymous identifier labels each step. Previous studies diverge in their findings about how much evidence of user interactions (Q and C) should be used for feedback: Tan et al. report in [5] that long term search history may improve web retrieval, while the authors of [4] argue to emphasize short-term query context. Also, Gong et al. question whether clicked data should be accepted as positive evidence of a document being relevant without a quality

[1] A European news agency: http://picture.belga.be/picture-home/index.html, log data collected within the VITALAS project: http://vitalas.ercim.org.

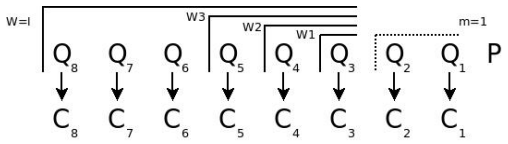

Fig. 1. Descriptive parameters of a search session: length $L = 8$, current observation at step $l = 6$, gap to a purchase $m = 1$ and history window $W = 1, 2, 3, max$.

metric [2], while Joachims et al. report on user studies where the quality of implicit feedback from clicks can compete with explicit judgements [3], especially when the additional burden on users in providing explicit feedback is taken into account.

In the pilot study described here, we consider a scenario where the retrieval system is expected to take advantage of the recorded query session, in an attempt to rank the user's purchases on top, early on in the session. This task is not trivial, as in any practical setting, the system does not know the total length L of the session. We define the *observation gap* m as the number of steps between the current interaction, *observation step* l, and the actual purchase P. Like session length L, m is not observable at state l. The open parameter that the system can choose is the size of query history window W. Fig. 1 summarizes our view on query logs, and the notation used in the paper.

2 Adding clicked documents as additional query terms

In [1], Balog et al. describe a series of experiments that compare the effectiveness of various language modelling approaches that exploit query expansion from explicit user feedback. The original formulation in [1] explores different assumptions about the cognitive process of selecting a set of relevant documents for feedback: they are taken to be grasped by a variation on a two steps generative process. When we apply this model to our scenario, viewing clicks as if they were explicit relevance assessments, the best performing setting according to the evaluation of [1] would first select a picture annotation from a set of clicks with probability $P(d|C)$ and subsequently pick a term from that annotation with probability $P(t|d)$. We follow [1] in not making any additional hypothesis about the dependence between queries and clicks, hence the click probability is uniform and the probability of finding a term t in the clicked annotation will be

$$P(t|C) = \sum_{d \in C} P(t|d) \cdot P(d|C) = \frac{\sum_{d \in C} P(t|d)}{|C|}. \tag{1}$$

Term frequency $\#(\cdot)$ in an unsmoothed maximum likelihood estimate is a crude measure of term importance $P(t|d)$ within a document, yielding

$$P(t|d) = P_{ML}(t|d) = \frac{\#(t,d)}{\sum_{\tau \in d} \#(\tau, d)}. \tag{2}$$

In this setting, using the top K terms with highest probability $P(t|C)$ to formulate an expanded query simply corresponds with using the K most frequent terms from each clicked document, considering each click equally important. In our case however, due to the relatively short picture annotations, after excluding stop words even for small K a large part of the expansion terms will be chosen among terms that appear just once in the document: we need an additional hypothesis to 'break the ties' may the most frequent terms be exhausted.

Qualitatively, we noticed how users formulate a query mostly based on previously examined documents. We make therefore the additional hypothesis that term importance also depends on the degree of surprise that a user experiences when reading the annotation, and discriminate between unique terms based on an entropy metric. We expand then a query Q into a new query $\hat{Q}$ with the K most frequent terms in clicked annotations weighted by the value of their γ-encoding for the entropy of the term distribution; since γ-encoding is prefix free, none of the document's terms will have exactly the same weighted frequency.

3 Evaluation

For our experiments on the Belga data, we have assumed session boundaries whenever the period of inactivity between two successive actions exceeded a 15-minute timeout. Queries are defined as the complete strings that are submitted in the search box, and split into query terms considering whitespace as delimiter. The Belga data contain 1003 sessions that consist of 3 to 13 steps before a purchase is observed, the subset that we use in our experiments. The distribution of sessions over session length L is given in Table 1.

Table 1. Distribution of sessions versus session length L.

Session Length	L=3	L=4	L=5	L=6	L=7	L=8	L=9	L=10	L=11	L=12	L=13
# Sessions	210	184	150	113	110	64	56	36	42	23	15

Using Lemur[2] out of the box, we retrieve 1000 images from the Belga collection and estimate the effectiveness of ranking the purchased pictures per session. As some sessions have recorded multiple purchases, we opted to report on Mean Average Precision (MAP), but the trends in the results are identical when reporting Mean Reciprocal Rank (MRR).

We compare three methods. The baseline method (Q) simply considers as a query the union of all the query terms in window W. The blind relevance feedback method (BRF) expands the query constructed as in the baseline method with the top 3 expansion terms from the top 5 ranked documents. The final method (Q+C) applies the query expansion approach described in the previous section, expanding the query produced in the baseline method with the $K = 5$ expansion terms derived from each of the clicked documents.

[2] `http://www.lemurproject.org/`

We simulate a system that operates under real conditions: i.e., with unknown session length L (and thus unknown observation gap m), attempting to improve the ranking of the purchased images. We vary window length ($W \in \{1, 2, 3, l\}$), and compare the success of ranking the purchases with the three methods described.

Our evaluation aims to single out the effect of the unobservable m while assessing the dependence on the history window parameter W. We first aggregate the performance over all observation gaps m (results summarized in Table 2).

Table 2. Average MAP over m: the MAP at $W = 1, m = 0$ is 0.0578.

MAP	Q only	Q + C	BRF
$W = 1$	0.0114	0.0448	0.0106
$W = 2$	0.0144	0.0526	0.0127
$W = 3$	0.0144	0.0573	0.0134
$W = l$	0.0114	0.0483	0.0104

Rank bias turns out hard to overcome in these experiments based on query logs: the MAP of 0.0578 obtained by a system not using any feedback ($W = 1, m = 0$), is superior to all other settings. If we however look into the settings where we 'look ahead' (the observation gap $m > 0$), then we conclude that a short-term context gives the best performance.

Next, we investigate the performance of the system as function of the distance to a purchase, considering a fixed observation gap and averaging only onto different session lengths. The top part of Fig. 2 shows that the baseline effectiveness (using past queries only) mainly depends on the distance to P, irrespective of the amount of history taken into account. Blind relevance feedback leads to slightly worse results. The bottom part of Fig. 2 demonstrates how click history could be a useful source of positive relevance feedback: more click history ranks the true purchases higher, with results for $W = l$ more than twice as effective as the setting ignoring session history ($W = 1$). We also see that using clicks reduces the dependence of effectiveness obtained upon the unobservable m. (Unfortunately, with the exception of the previously mentioned case of $m = 0$; which we attribute to rank bias.)

4 Conclusions and future work

We have investigated the effectiveness of using QL data to improve the retrieval performance, and the quality of clicked data as a source of implicit relevance feedback. The preliminary results obtained show that information about previous user interactions with the system may help improve overall performance. Predicting a purchase early in the session remains difficult, but taking a moderate amount of session information into account seems beneficial. Our conclusions are preliminary, as they are based on relatively straightforward methods - we

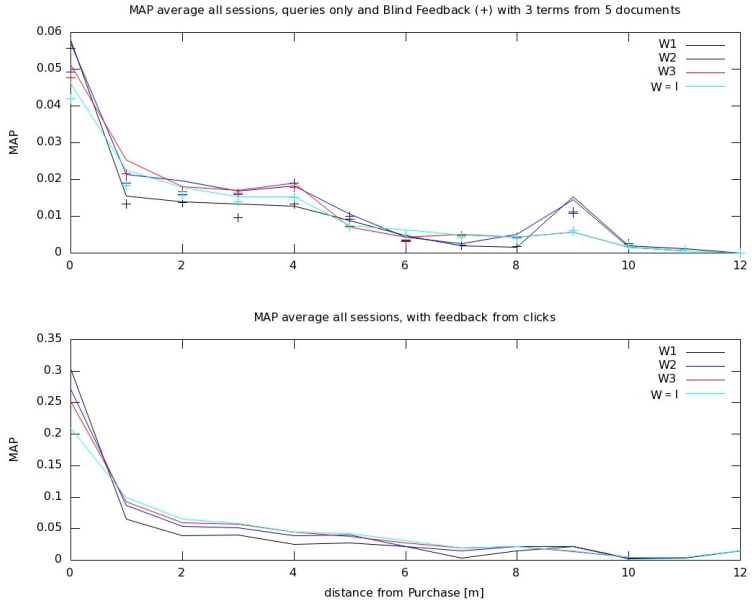

Fig. 2. MAP for $W = 1, 2, 3, l$ with (bottom; Q+C) and without (top; Q) additional terms from the clicked data; (+) in the top figure plots BRF results.

have not weighted query terms, and did not tune all parameters of the retrieval model to the collection. Apart from exploring better methods for query representation, we would like to relax the assumption that a session relates in its entirety to a single, static information need, and investigate in more depth the relation between the documents visited and the subsequent user queries issued.

References

1. K. Balog, W. Weerkamp, and M. de Rijke, "A few examples go a long way: constructing query models from elaborate query formulations," in *SIGIR*, 2008.
2. B. Gong, B. Peng, and X. Li, "A personalized re-ranking algorithm based on relevance feedback," in *Advances in Web and Network Technologies, and Information Management*, 2007, vol. 4537, pp. 255–263.
3. T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in *SIGIR*, 2005.
4. G. Pandey and J. Luxenburger, "Exploiting session context for information retrieval - a comparative study," in *Advances in Information Retrieval*. Springer Berlin / Heidelberg, 2008, vol. 4956, pp. 652–657.
5. B. Tan, X. Shen, and C. Zhai, "Mining long-term search history to improve search accuracy," in *Proceedings of SIGKDD*, 2006, pp. 718–723.