Using query logs and click data to create improved document descriptions

Maarten van der Heijden* mvdheijd@sci.ru.nl
Max Hinne* mhinne@sci.ru.nl
Wessel Kraaij* w.kraaij@cs.ru.nl
Suzan Verberne† s.verberne@let.ru.nl
Theo van der Weide* tvdw@cs.ru.nl

ABSTRACT

Logfiles of search engines are a promising resource for data mining, since they provide raw data associated to users and web documents. In this paper we focus on the latter aspect and explore how the information in logfiles could be used to improve document descriptions. A pilot experiment demonstrated that document descriptors extracted from the queries that are associated with documents by clicks provide useful semantic information about documents in addition to document descriptors extracted from the full text of the web pages.

1. INTRODUCTION

Search engine query logs and associated click-through data have become a very important resource to improve document ranking [8, 14, 9, 11]. Since click-through data can be used as a direct measure of the relevance of web pages for queries, and a large proportion of the queries are frequent, it is possible to re-rank result lists based on these observations. This has led to an increased interest of machine learning methods applied to ranking. However, since it has been shown that it is relatively easy to reveal the identity of searchers using search engine query logs and click-through data, search engine companies have become very reluctant to make these data-sets available for research. Recently a data set consisting of 12 million query and click pairs was made available by Microsoft for the purpose of scientific study.

Unfortunately, the data-set does not include the full information that was available to the user when he clicked (web page title and snippet). This means that it is not possible to identify the effect of the document summaries (snippets) that the users saw for pages they clicked on. Since the query log set is a few years old and re-crawling the data would yield inaccurate results, we have decided to focus our attention on exploring what latent (semantic) information is available and how it could be applied.

*Dept. of Computer Science, Radboud University Nijmegen
†Dept. of Linguistics, Radboud University Nijmegen

WSCD '09, Feb 9, 2009 Barcelona, Spain.

An initial investigation of the data set showed that the assumption that one session would consist of just one query topic was wrong. We therefore decided to concentrate our attention on the URLs and their associated query clouds (the weighted set of query terms that led to a URL) and defer study of query reformulation models.

The focus of our study is to explore whether (semantic) associations between terms can be mined from the data set. We investigate whether these associations can be approximated both by a representation based on full text of documents, and by a representation based on query data. We will initially model these semantic associations as simple unigram association models (probabilistic lexical models). These association models could be applied in several ways:

- Suggesting related terms that can be used for query reformulation or query expansion.
- Creating an expanded document language model, which could be used to improve a method for clustering web pages.
- Generating snippets (in the form of query term clouds) for web pages based on the associated and re-formulated queries.

For the current experiments, a future objective is to provide better snippets (summaries) for the documents retrieved by the search engine. The snippets that we aim at are not only based on the contents of the document and the query terms that were inserted by the current user (such as in the state of the art search engines), but also on the query terms from other users that have led to the document in previous search sessions.

The remainder of this paper is organized as follows. Section 2 discusses some related work to mining semantic associations of document and query terms. Section 3 describes the experiment we conducted. The results are presented and discussed in Section 4. Lastly Section 5 provides our conclusions and suggestions for further research.

2. RELATED WORK

Despite the fact that few public query log data sets are available for research, this type of data has spawned several new IR research subcommunities. The most visible of these communities explores the use of click-through data as a training set for machine learning approaches for ranking (learning to rank[1]). Example studies include: search

[1]http://research.microsoft.com/users/LETOR/

engine optimization using click-through data [8], treating click-through data as implicit feedback [9], optimizing web search using web click-through data [14] and active exploration for learning rankings from click-through data [11]. Click-through data is also a good resource for query reformulation or query expansion [4, 5, 7, 10].

Our work aims at enriching document descriptions by mining the queries that are linked to documents through click events. Related work is the study by Baeza Yates and Tiberi [1] where queries are considered as social tags. The click-through data can be modeled as a graph, where queries are nodes in the graph and are connected when they share clicked URLs. Just like folksonomies, semantic relations can be mined from this graph. Other related work is summarization based on click-through data. Here, query terms linked to a document help to improve summarization. Sun et al [12] describe two possible methods to incorporate click data. Query terms that led to a click are probably important and can thus be given more weight when determining the significant words that are important features for sentence extraction. Similarly, term weight can be adapted in LSA-based summarization. Boydell and Smyth [2] show that social bookmarks (e.g. del.icio.us) can be used as pseudo queries, to harvest snippets that in turn can be used for 'social summaries'. Both studies show that queries and bookmarks are valuable sources of condensed content descriptions.

3. EXPERIMENTS

The objective of our experiments is to investigate the quality of a (shallow) semantic[2] representation of a web document based on query terms as an addition to more common semantic representations based on full document text (such as in the text snippets presented in search engine's result lists). In our evaluation, we compare the representation based on query terms to the representation based on the full document text in order to see whether the query-based representation is a valuable addition to the document-based representation.

3.1 Data

The Web Search Click Data (WSCD) consists of approximately 15 million queries from US users entered into the Microsoft Live search engine. It is also referred to as the Microsoft 2006 RFP dataset. The data was collected over a period of one month in the spring of 2006. For each query the following details are available: a query ID, the query itself, the user session ID, a time-stamp, the URL of the clicked document, the rank of that URL and the number of results. For privacy reasons, all email addresses have been replace by an underscore (_) and long number sequences by compound (#).

Since the WSCD collection only contains the url of each of the documents that was clicked on, and not the contents of the documents, we first created a full-text document collection for a subset of the click-through data in order to be able to do analyses on full document texts. For this purpose, we used the Wikipedia XML document collection as it was used in INEX 2006 [6]. This collection was created in the same year as the WSCD click-through data, and covers most Wikipedia urls from the Microsoft RFP 2006 dataset. Although Wikipedia documents are dynamic documents, we assume that the pages crawled in 2006 are a more accurate match with the pages at the time they were clicked on by MS Live users than the current versions of those pages.

From the WSCD click-through data, we extracted all urls that contain the substring *http://en.wikipedia.org/wiki*. For the first 20,000 Wikipedia urls, we retrieved the complete documents from the INEX corpus using the Wumpus engine [3] with the document titles as input[3]. Due to duplicate urls and mismatches between the corpus and the url set, this resulted in a subcorpus of 9,575 Wikipedia documents with 31,475 queries that have led to these documents. The median of the number of clicks per document was 1 and the mean number of clicks 2.45.

3.2 Method: building the representations

For the development of the representations, we use the following resources: all queries from the WSCD click-through data, an Indri index of the complete INEX 2006 Wikipedia corpus created by Lemur[4], the textual content of our subset of 9,575 Wikipedia documents and the queries that led to these documents. In the current experiments, we chose $tf.idf$ as function for weighting the terms in the models. We built the following two data structures for each Wikipedia document:

1. A $tf.idf$ term vector based on the full document text. Term frequency (tf) is the number of occurrences of the term t in the document. Document frequency is the number of documents in the complete Wikipedia 2006 corpus in which t occurs. *idf* was determined as $\ln(|D|/df(t))$. We get document frequency for each term in the corpus by applying the *dumpindex* function from Lemur to the complete INEX 2006 index, in which $|D|$ is total number of documents in the Wikipedia corpus.

2. A $tf.idf$ term vector based on the query terms that led to the document. Term frequency (tf) is the number of occurrences of the term t in the set of queries that have led to the document. Document frequency is the number of queries in the complete WSCD click-through data in which t occurs. *idf* was determined as $\ln(|D|/df(t))$, in which $|D|$ is total number of queries in the click-through data.

3.3 Evaluating the representations

We evaluated the two approaches with human assessments of condensed versions of the semantic representations. To this end, we extracted a selection of 29 Wikipedia articles from our sub-corpus of 9,575 documents[5]. The list of selected documents (topics) can be found in Table 1. Subsequently the two different term vector representations (one based on full text and one based on query terms) were constructed for each document. Finally, the top 10 most identifying terms were extracted from each representation by ranking the terms according to $tf.idf$. The choice of $tf.idf$

[2]The intended interpretation is 'lexical semantics', a representation of the main concepts in a document.

[3]We used the Wikipedia-specific feature that urls contain the exact title.

[4]www.lemurproject.org

[5]Our selection was largely random, but with the requirement that the selected documents cover general knowledge topics.

Table 1: 29 Wikipedia articles (topics) selected from the sub-corpus of 9,575 Wikipedia articles for manual assessment

Baroque	Lasagna	Nasal cavity	Woodworking joints
Baseball	Linux	Nostradamus	World War I
Batman	Madonna (entertainer)	Plato	Yahtzee
George W. Bush	Martin Luther King Jr.	Postage stamp	Zen
China	Milk	RFID	Zimbabwe
Colorblindness	Moon	Rock and Roll	Zorro
Earthquake	Mount Everest	Roman Empire	
Eiffeltower	Wolfgang Amadeus Mozart	The Wonderful Wizardof Oz	

for ranking the document terms implies that these terms are expected to be relatively specific (due to the heavy *idf* weight) to the topic.

For each Wikipedia topic, these two lists of terms were merged into an alphabetic list. On average, 14 terms are listed per topic, indicating a considerable overlap between the two shortlists of 10 terms. Twelve human assessors were given six topics each together with the alphabetic list of terms for each topic. Each topic was assessed by two or three assessors. They were asked to label minimally three, maximally eight terms that are (according to them) most strongly associated with the topic. If an assessor was not familiar with the topic, she could use the full text of the Wikipedia article as a reference.

The assessments were processed as follows: each vote for a term is counted as a positive vote for the representation where the term occurs in. Votes for terms that occur in both representations count for both representations. An example of the human assessments is shown in Table 2.

Table 2: Assessments for one example topic: 'Plato'. A1 is assessor 1; A2 is assessor 2. An 'x' means that the assessor judges the term as one of the terms that are most strongly associated to the topic.

Terms	A1	A2
dialogues	x	x
forms		
intelligible		x
knowledge	x	
perceptual		
philosophy	x	x
plato	x	x
platonic		x
platos		
republic		x
see		
socrates	x	x
works		

We used the processed assessments to evaluate how the two representations (full-text or query-based) were judged by the assessors. We not only counted the votes for each of the representations, but also used the term rank (based on $tf.idf$), representing the importance of the term in the representation of the Wikipedia page. We applied the following scoring function per topic:

$$S(R,T) = \sum_{t \in L} \frac{votes(t)}{rank(R,t)}, \quad (1)$$

in which $R \in \{R_q, R_f\}$ is either the query or the full text representation, T the title of the current document (topic), L the list of presented terms, $votes(t)$ the number of votes for a given term t and $rank(R,t)$ the $tf.idf$ rank of t in that representation. This is essentially a mean reciprocal rank function, where terms that have been selected by both assessors are favored.

4. RESULTS AND DISCUSSION

On average, the assessors selected 5 terms per topic. The inter-annotator agreement for this task was moderate ($\kappa = 0.58$)[6]. The moderate agreement is mainly due to a relatively large range of number of terms to be selected (3–8) and the presence of form variants in the same topic (e.g. both *batter* and *batters* for the topic *baseball*). Nonetheless, using (1) we obtained interesting results.

Let us first illustrate our results with an example. For the Plato example (Table 2), calculating the scores for both representations resulted in $S(R_q, \text{'Plato'}) = 3.88$ and $S(R_f, \text{'Plato'}) = 2.73$. This indicates that for this document, the assessors preferred the terms generated by the query representation over those generated by the full text representation.

Overall, we found that for 18 out of 29, the assessors judged the terms extracted from the query term representation to be more salient than those extracted from the full-text representation. For the other 9 topics, the terms extracted from the full-text representation were judged more salient.

Although we experimented with only a small set of documents and assessments, our results suggest that terms extracted from a representation based on query term information provides useful additional semantic information about documents. Our experiment show that click-through associated query text can be used as a valuable semantic annotation to be used in search engine snippets additional to a semantic description based on the text of the document.

We performed an analysis of the salient terms that were extracted from both representations and the terms that were selected by the assessors to see what the differences are. Consider Table 3. For the topic *Zen*, assessors selected the terms *buddhist*, *meditation*, *zen* and *practice*.

We suggest that the differences between the salient terms from the two types of representations are due to the type

[6] $\kappa = \frac{P(A)-P(E)}{1-P(E)}$, with $P(A)$ the observed agreement between the assessors (simple matching), and $P(E)$ the chance agreement. Since $P(E)$ is generally difficult to calculate exactly, we estimated its value as the probability that two assessors randomly vote for the same term (≈ 0.55). The terms are considered independent in this regard. From the data we observed $P(A) = 0.81$, hence $\kappa = \frac{0.81-0.55}{1-0.55} = 0.58$.

Table 3: Salient terms for the document *Zen* based on $tf.idf$ as suggested by R_q and R_f, ranked in descending order. An asterisk (*) indicates that the term was selected by at least one of the two assessors.

R_q	R_f
zen*	zen*
practice*	koan
meditation*	teachers
zazen	zazen
teacher	lineage
japanese	meditation*
buddhist*	soto
well	practice*
centers	rinzai
life	dharma

of identifying terms the representations provide. Terms extracted from R_q are usually terms that often co-occur with the topic, whereas terms extracted from R_f are more explanatory and contain WordNet-like associations such as 'is-a' and 'has-a'. This is because the full-text representation originates from a set of encyclopedic (Wikipedia) documents, while the query term representation can be seen as a collection of user-generated annotations for these documents. A second difference is that the terms extracted from R_f are relatively specific to the topic, whereas the terms extracted from R_q are more general, since a user searching for information on a topic is expected to use less specific terms than the author of the Wikipedia page on the topic.

5. CONCLUSION AND FUTURE WORK

In this paper we investigated the possibility of mining semantic relations from search engine query logs. Query terms that led to a click on a document imply some semantic relation between the term and the document. Our experiment with human assessors seems to support this idea, although the sample size was too small to draw more than tentative conclusions.

In future work we intend to further explore the potential of click-through data by using more complex representations (bi-gram or bi-term instead of unigram) to find word associations. Moreover, we plan to experiment with pointwise Kullback-Leibler divergence [13] or pointwise cross entropy as alternatives to TFIDF for ranking terms by their importance.

Finally, we aim to implement a retrieval interface in which the snippets of the retrieved documents are extended with terms extracted from click-through query representations such as presented in this paper. We will set up a user experiment to find out whether these extended snippets lead to better motivated document clicks and thereby to better user satisfaction.

6. REFERENCES

[1] R. Baeza-Yates and A. Tiberi. Extracting semantic relations from query logs. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 76–85, New York, NY, USA, 2007. ACM.

[2] O. Boydell and B. Smyth. From social bookmarking to social summarization: an experiment in community-based summary generation. In *IUI '07: Proceedings of the 12th international conference on Intelligent user interfaces*, pages 42–51, New York, NY, USA, 2007. ACM.

[3] S. Buttcher, C. Clarke, and G. Cormack. Domain-specific synonym expansion and validation for biomedical information retrieval (multitext experiments for trec 2004). 2004.

[4] H. Cui, J. Wen, J. Nie, and W. Ma. Probabilistic query expansion using query logs. In *Proceedings of the 11th international conference on World Wide Web*, pages 325–332. ACM Press New York, NY, USA, 2002.

[5] H. Daume and E. Brill. Web search intent induction via automatic query reformulation. In *Human Language Technology Conference/North American chapter of the Association for Computational Linguistics (HTL/NAACL)*, 2004.

[6] L. Denoyer and P. Gallinari. The Wikipedia XML corpus. *ACM SIGIR Forum*, 40(1):64–69, 2006.

[7] C. Huang, L. Chien, and Y. Oyang. Relevant term suggestion in interactive web search based on contextual information in query session logs. *JASTIS*, 54(7):638–649, 2003.

[8] T. Joachims. Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, New York, NY, USA, 2002. ACM.

[9] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference*, pages 154–161. ACM New York, NY, USA, 2005.

[10] J. Mostert and V. Hollink. Effects of Goal-Oriented Search Suggestions. In *Proceedings of the Belgian-Dutch Conference on Artificial Intelligence (BNAIC 2008)*, 2008.

[11] F. Radlinski and T. Joachims. Active exploration for learning rankings from clickthrough data. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 570–579. ACM Press New York, NY, USA, 2007.

[12] J. Sun, D. Shen, H. Zeng, Q. Yang, Y. Lu, and Z. Chen. Web-page summarization using clickthrough data. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 194–201. ACM New York, NY, USA, 2005.

[13] T. Tomokiyo and M. Hurst. A Language Model Approach to Keyphrase Extraction. In *Proceedings of the ACL Workshop on Multiword Expressions*, pages 34–40, 2003.

[14] G. Xue, H. Zeng, Z. Chen, Y. Yu, W. Ma, W. Xi, and W. Fan. Optimizing web search using web click-through data. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 118–126. ACM New York, NY, USA, 2004.