Session Segmentation Based on Document Metadata

Tomáš Kramár*
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 3, 842 16 Bratislava, Slovakia
kramar@fiit.stuba.sk

Abstract

It has been shown that the search personalization can greatly benefit from exploiting user's short-term context – his immediate needs and focus. But to achieve that, we need to be able to tell when the context changes; we need to be able to divide the user's activity into segments, where each segment captures user's single goal and focus. Many different approaches exist, but their major weakness is that they build inaccurate models that do not include user's implicit feedback. We present a method for segmenting queries into search sessions which is based on document metadata and incorporates implicit feedback and as such is able to build more accurate context model.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Query formulation, Selection process, Search process

Keywords

personalization, search, short-term contexts, personalized search

1. Introduction

Search engines play a crucial role in accessing the amount of content on the Web. Users interact with search engines by entering few keywords, which describe their intent and expect the machine to provide a list of relevant documents. But this model has several known disadvantages:

- the number of keywords is usually low, typically 1-3 keywords [9];
- many of the words are ambiguous; imagine a word "jaguar" which can refer to an animal, a car and even has less-known meanings such as a game console or German battle tank;
- the queries are almost never accurate [5], they are either too generic or too specific, but almost never exactly aligned with the specific intent the user has in mind.

When we combine the impact of each of the described problem, we come to a conclusion, that finding the relevant document is indeed a difficult task, both for the user and the search engine.

To mitigate this problem, several approaches to search personalization have been researched [2, 1], each with the ultimate aim to help the user find the relevant content, without trying to change how humans think, or work [11]. However, most of the personalized search methods only concetrate on the long-term interests, completely ignoring the short-term goals, which are at least equally important. But to be able to use short-term goals a personalization system must know the exact moment the user changes her intent, so that it can start and use a new context. The task of detecting this change is referred to as search session detection (segmentation). The term *search session* was never formally defined in the literature and its meaning differs in the different works. In this work, we assume that search session is a sequence of search related actions with the single underlying informational intent.

2. Related works

The most widely used approach to search session segmentation is to compare temporal distance of the queries. If two queries were issued with a time difference larger than a predefined threshold, it is assumed that a new session started, and the existing session is split at that place. This technique was first described in [3], establishing the threshold of 25,5 minutes. Due to its simplicity both in concept and implementation, this technique is widely used and there are modification which differ in the cutoff size. The temporal-based approach has the disadvantage, that it is unable to detect sessions, which are split within a short time. The long time between searches may not mean that the user's interest changed and vice versa, two consecutive queries might bear very different search intents.

Another approach uses lexical distance of the queries. This idea compares content of two queries to detect if the

*Doctoral degree study programme in field Software Engineering. Supervisor: Professor Mária Bieliková, Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies, STU in Bratislava. Work described in this paper was presented at the 7th Student Research Conference in Informatics and Information Technologies IIT.SRC 2011 and awarded by Dean's Award.

Kramár, T. Session Segmentation Based on Document Metadata. Information Sciences and Technologies Bulletin of the ACM Slovakia, Special Section on Student Research in Informatics and Information Technologies, Vol. 3, No. 2 (2011) 64-66

intent has changed. Example of this approach is stated in [8]. The main drawback of this method is that it leads to a high amount of false positives. There are many instances where two queries are completely dissimilar (they share no common words), yet their underlying intent is the same. Consider a user searching for *"IR"* and *"information retrieval"* afterwards. Using the lexical distance approach would incorrectly yield two separate sessions.

Method described in [6] combines both temporal and lexical distances. They use vector-space representation, where each pair of following queries is placed in the space. If the query pair fits into the space bounded by the subplane delimited by two edge cases (i) two parallel, but dissimilar queries, (ii) the same queries, but executed long time apart, it is considered an extension of the current session. Although this combination can achieve better results than each of the methods alone, it is still prone to the aforementioned problems.

As stated, related queries do not provide sufficient data to match similar intents, but there are approaches that use signals from the retrieved documents. In [10], authors used vectors of titles and snippets of 50 top-ranked documents for the query. The session is split as determined by comparing cosine similarities of two following queries. Another approach uses document keywords. In [4], authors extracts document's keywords using the TF.IDF scheme and map them to ODP categories. The session is then split when the ODP category changes. This approach however fails to account for user's real interests. When the user enters a query and clicks and views some documents only to realize that the results are wrong and the query needs to be reformulated, this approach has already used these faulty results to decide upon session segmentation.

Our work extends and differs from existing research in several important ways. First, we only consider the documents that the user deemed useful. To measure the usefulness, we partially rely only on the clickthrough data and only consider the search results that the user actually viewed, but we also use implicit feedback indicators to weight the contribution of each of the documents. The search sessions are segmented by comparing their metadata (both machine and human extracted), and to increase the chance of a successful match, we extend the metadata with ConceptNet relations.

3. Segmenting search sessions

Our method of search session segmentation operates in the following steps:

1. For each query, find the candidate set of useful pages – the documents that were clicked from the search results page (SERP). We use the value of *referrer* – each Web page that was followed from the SERP would have the SERP URL address as referrer.

2. Prune all documents that were viewed (their *time on page* attribute is) less than 4 seconds. The value of this parameter was selected arbitrarily, based on the manual analysis of the logs. We assume that any document displayed for less than 4 seconds did not interest the user, so it is not contributing its features to the segmentation algorithm.

3. If there is a query which has no document views associated with it (either there was no SERP click, or all SERP clicks were pruned by the 4 seconds window) then its metadata are initialized to an empty set.

4. For all other queries, their respective metadata sets are built by unioning all metadata from all associated documents.

After the final step, each query has associated a set metadata which describes the search results that the user found interesting. We also build a second set of enhanced metadata for each query by querying the ConceptNet [7] API for related terms.

4. Evaluation

To preliminary evaluate our approach to search session segmentations, we used a small log collected on the PeWeProxy[1] platform, consisting of 245 searches collected over the period of 4 days. To obtain a baseline to compare against, these queries were manually segmented into search sessions by a human evaluator. To remove a possible time bias, the segmentation user interface did not contain information about the time of query – only the text of the query itself.

Table 1 summarizes the results obtained with various segmentation methods. Approach in the table refers to the specific method that was used to segment the logs. In *lexical* approach, we used query similarities – two following queries sharing at least one common keyword were considered to be a part of the same session. *Temporal* approach refers to time based similarity. The number in the name indicates the cutoff used. *Temporal30m* means that queries were segmented by using a 30 minutes inactivity cutoff time, *temporal5m* uses a 5 minute cutoff. In *metadata* we compared similarity between the sets of query metadata – for two following queries, if the intersection of their metadata sets was non-empty, the queries were considered part of the same session. Similarly *enhanced_metadata* compares sets of ConceptNet enhanced metadata.

To evaluate the quality of each approach, sessions were compared to the baseline (manually created sessions) and precision and recall were calculated using the following methodology:

Precision indicates the internal coherence of the session. First, queries from automatically detected session are linked to the manually detected sessions and precision is calculated as the ratio of the cardinality of the best match to the cardinality of the whole session. (e.g. the automatic method detected a session of queries A, B, C, D and these queries are linked to sessions 1, 1, 1, 2 in the manually created sessions. The precision in this case is 3/4.

Recall indicates the completeness of the session. It is calculated as the ratio of cardinality of the best match against the manually created sessions to the cardinality of the best-matched session. In the above example, if the session 1 contains 7 queries, the recall is 3/7.

[1] PeWeProxy, `http://peweproxy.fiit.stuba.sk`

Table 1: Results of search session segmentation using various approaches

approach	avg(precision)	std	avg(recall)	std	f-score
lexical	0.985	0.078	0.499	0.410	0.662
temporal30m	0.932	0.153	0.489	0.380	0.641
temporal5m	0.974	0.101	0.346	0.373	0.510
metadata	0.992	0.055	0.276	0.312	0.432
enhanced_metadata	0.932	0.167	0.272	0.284	0.421
lexical+metadata	0.969	0.108	0.540	0.411	0.694
lexical+enhanced_metadata	0.887	0.190	0.563	0.396	0.689

We see that when using the classical methods, the lexical approach outperforms temporal methods, with the widely used 30 minute cutoff yielding slightly better performance. Metadata based approach did not outperform temporal approaches and yielded low recall. This is due to the fact that the extracted metadata generally covers wide area and metadata from similar pages on the same topic do not necessarily overlap.

We expected that enhancing the metadata would improve performance by connecting similar words and improving the chance for a match, yet the performance of the enhanced metadata was worse than that of the metadata alone. This is rather surprising, but after closer inspection of the metadata and ConceptNet connections, we believe that this was caused by several factors: (i) queries were generated by technical users and the majority of them deals with specific technologies and problems, which are not present in the ConceptNet database – the majority of metadata was thus left unexpanded; (ii) in many cases, the ConceptNet relations are too general and at very different conceptual levels (e.g. the term "jaguar" is linked to term "black") – in turn, the query metadata was enhanced with many generic and abstract terms, leading to a false positive matches thus reducing precision and recall.

On closer inspection of the results, we noticed that the metadata based approach excels at linking queries where lexical similarity fails. We combined the two approaches, and employed the metadata based similarity only when the lexical similarity did not find a match. This approach yielded the best score.

5. Conclusions and future work

In this paper, we presented our approach to search session segmentation, which calculates query similarity by comparing similarities of the metadata extracted from the documents clicked from the search results page of the respective query. We only consider documents that the user clicked and that she found useful. We evaluate the usefulness of the page by the fact that it was clicked within a search session and by employing the implicit feedback signals, namely the time spent on page. Using this approach, we are able to overcome the disadvantages of using metadata from all documents returned as a response to the query – we discard badly formulated queries or misleading search results.

In the next work, we will focus on building a larger dataset of manually segmented sessions. We plan to use the personalized proxy server platform to allow users themselfs to segment their own sessions as they search. We also plan on evaluating other means of enhancing the document metadata, e.g. by using automatically generated hypernyms.

Acknowledgements.
This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11 , and it is the partial result of the Research & Development Operational Programme for the project Research of methods for acquisition, analysis and personalized conveying of information and knowledge, ITMS 26240220039, co-funded by the ERDF.

References

[1] A. Andrejko. Novel approaches to acquisition and maintenance of user model. *Inf. Sciences and Tech. Bulletin of the ACM Slovakia*, 1:1–10, 2009.

[2] M. Barla. Towards social-based user modeling and personalization. *Inf. Sciences and Tech. Bulletin of the ACM Slovakia*, 3:52–60, 2011.

[3] L. D. Catledge and J. E. Pitkow. Characterizing browsing strategies in the world-wide web. *Comput. Netw. ISDN Syst.*, 27:1065–1073, 1995.

[4] M. Daoud, M. Boughanem, and L. Tamine-Lechani. Detecting session boundaries to personalize search using a conceptual user context. In S.-I. Ao and L. Gelman, editors, *Advances in Electrical Engineering and Computational Science*, volume 39 of *Lecture Notes in Electrical Engineering*, pages 471–482. Springer Netherlands, 2009.

[5] D. Downey, S. Dumais, D. Liebling, and E. Horvitz. Understanding the relationship between searchers' queries and information goals. In *Proc. of the 17th ACM Conf. on Inf. and Knowledge Management*, CIKM '08, pages 449–458. ACM, 2008.

[6] D. Gayo-Avello. A survey on session detection methods in query logs and a proposal for future evaluation. *Inf. Sci.*, 179:1822–1843, 2009.

[7] C. Havasi, R. Speer, and J. Alonso. ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge. In *Recent Advances in Natural Language Processing*, 2007.

[8] B. J. Jansen, A. Spink, C. Blakely, and S. Koshman. Defining a session on web search engines: Research articles. *J. Am. Soc. Inf. Sci. Technol.*, 58:862–871, 2007.

[9] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf. Process. Manage.*, 36:207–227, 2000.

[10] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. In *Proc. of the 14th ACM Int. Conf. on Information and Knowledge Management*, CIKM '05, pages 824–831. ACM, 2005.

[11] V. Vaneková. Preferential querying for the semantic web. *Inf. Sciences and Tech. Bulletin of the ACM Slovakia*, 2:137–148, 2010.