Activity-based Search Session Segmentation

Samuel Molnár*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
xmolnars1@fiit.stuba.sk

Automatic search goal identification is an important feature of personalized search engine. The knowledge of search goal and all queries supporting it helps the engine to understand our query and adjust sorting of relevant web pages or other documents according to our current information need. To improve the goal identification the engine uses other factors of user's search context and combine them together by different relevance weight. Although, most of factors utilized for goal identification involve only lexical analysis of user's queries and time windows represented as short periods of user's inactivity.

Recent works tackle the problem of search session segmentation by lexical, semantic and behaviour driven approaches [1, 2, 3]. Most of the approaches are based on utilization of time windows within defined interval. Jones and Klinkner [2] identified the task identification problem as a supervised classification problem, and tried four different timeouts (e.g. 5, 30, 60 and 120 minutes). Time-based approaches lack the precision in segmenting search session [4], since many search task are interleaving and user's search intent might span multiple days. Therefore, time-based or temporal features (e.g. inter query threshold or thresholds of user's inactivity) are usually used in conjunction with lexical or semantic features.

Most of the proposed approaches consider lexical features of queries for determining query similarity [1, 2, 4, 5]. Jones and Klinkner [2] identified normalized levenshtein distance as the best lexical similarity measure for identifying goal boundaries. Semantic features of queries were utilized only in case of large corpora like Wikipedia[1] or Probase[2] [5, 6]. Proposed approaches propose 'wikification' for extending the meaning of a query in terms of concepts mined from Wikipedia or Probase.

---

* Supervisor: Tomáš Kramár, Institute of Informatics and Software Engineering
[1] Wikipedia: https://www.wikipedia.org/
[2] Probase: http://research.microsoft.com/en-us/projects/probase/

Spring 2014 PeWe Workshop, March 21, 2014, pp.13–14.

Similar approaches for segmenting search sessions focus on utilization of user's behaviour described as Markov model [7]. Markov model proposed by authors in [7] outperforms other methods based solely on lexical or semantic features.

In our work, we focus on utilizing user activity during search for extending existing lexical and time factors. By analysing user search activity such as clicks and dwell time on search results, we better understand which search results are relevant for user's current information need. Thus, we utilize user's implicit feedback to determine relevance between queries by search results they share. Strong relationships between queries provide similarity measure between queries by the number of shared link adjusted by user's implicit feedback. Semantic analysis of queries and search results snippets is another factor we introduced for clustering queries into sessions. Utilization of encyclopedias like Wikipedia and Freebase can provide a way of understanding concepts and user's intention behind the query and thus provide another clustering factor. We plan to integrate our model of weighted factors utilizing user activity and semantic analysis to existing search engines or servers like Elasticsearch.

Acknowledgement. This work was partially supported by the VEGA1 - Bielikova.

References

[1] OZERTEM, U. - CHAPELLE, O. Learning to Suggest : A Machine Learning Framework for. In SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. 2012. s. 25–34.

[2] JONES, R. - KLINKNER, K.L. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In Proceedings of the 17th ACM conference on Information and knowledge management. 2008. s. 699–708.

[3] CHEN, E. Context-Aware Query Suggestion by Mining Click-Through and Session Data. In Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '08 (2008). 2008. s. 875–883.

[4] JONES, R. - KLINKNER, K.L. Beyond the Session Timeout : Automatic Hierarchical Segmentation of Search Topics in Query Logs Categories and Subject Descriptors. In Proceedings of the 17th ACM conference on Information and knowledge management (2008). 2008. s. 699–708.

[5] HUA, W. et al. Identifying users' topical tasks in web search. In Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13. 2013. s. 93.

[6] LUCCHESE, C. et al. Identifying Task-based Sessions in Search Engine Query Logs Categories and Subject Descriptors. In Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11 (2011). s. 277–286.

[7] HASSAN, A. et al. Beyond DCG : User Behavior as a Predictor of a Successful Search. In. s. 221–230. .