SEARCH QUERY EXPANSION BASED ON SEARCHER'S ACTIVITY CONTEXT

Ľubomír VNENK, Mária BIELIKOVÁ

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
Lubomir.vnenk@zoho.com, maria.bielikova@stuba.sk

Abstract. Finding desired specific information feels often like looking for a needle in a bottle of hay. Success or failure of our search with current keyword based concept of the search still depends on how good and precise query we are able to specify. We follow in our research well known fact that search intentions correspond in most cases to the searcher's activity. We aim on carefully capturing searcher's activity within the scope of all applications to model activity context model as much precise as possible. We present an approach for analysis of connections between searcher's activity context and his actual search context to extend basic search context by information related to the actual search context. The more information about actual search context we have, the more precise keywords reflecting search intention we can find to extend the actual query.

1. Motivation

There is so many information on the Web that finding desired one is a difficult task. Number of various information is increasing each day and quality of information depends on many aspects. Different web-sites can also see information in different views and describe different piece of the information. Therefore to find a desired specific information query must be typed very precisely. However, user's query mostly consists in average of 2-3 keywords [6]. Few keywords cannot exactly express search intentions. It may be because of ambiguous query (keyword may has many meanings) or because of not specified desired information.

If the search is successful and the searcher finds the information he was looking for, it usually will not take longer than 5 minutes. However, if he is not successful, it can take more than 10 minutes to give up searching [4]. Moreover, the longer actual session lasts the more frustrated searcher will become [3]. Our goal is to make a search query more specific to find requested information faster to save searcher's time and to make his search session more

comfortable. To achieve this goal we proposed query expansion method based on searcher's activity context.

2. Query expansion method

To improve the state of the art we did not aim to enhance existing query expansion methods but we concentrated on providing much better inputs, i.e. activity context. The most of existing methods consider searcher's activity just in search-results page or in browser. We captured activity also outside web browser. Our method consists of three steps:

1. Capturing searcher's activity within the scope of his work on computer (also outside a browser).
2. Finding actual search extended activity context.
3. Expanding query by few most relevant keywords.

To capture searcher's activity within the scope of the whole computer we proposed and realized three independent loggers:

- *Tabber* as a desktop application logger developed in c# because of powerful libraries that provide an ability to hook each possible application.
- *Annota* as a Firefox browser logger of user's activity on each page the user visited (we used a component of the Annota system aimed at annotation and bookmarking of web resources [5], annota.fiit.stuba.sk).
- *Wordik* as a Microsoft Office Word logger made to log user's activity inside the application for the purpose of analyzing various styles and structure of documents user has been writing.

Loggers capture title of an application window each time the searcher has switched into the application so we were able to get precise click-through across all applications. We also capture copied/pasted text, because we consider this text as point of searcher's interest and important information.

We captured just raw data, so we needed to process them into searcher's activity context model. Each application has different purpose so we decided to model an activity context for each application, so the searcher's activity context model consists of many small applications' activity contexts. Logged data are captured in natural language, so we had to find out in which language the information was and apply NLP methods proposed for this language. Firstly, we filtered stop-words, because they have no meaningful value. Secondly, we stemmed the information to be able to find similarity between different inflected forms.

Second step of our method is to find actual search extended activity context. We hypothesize, that search activity context is the same as, or similar to, the activity context of an application that the searcher has recently used. It means that we should be able to find conjunction between a query and the related application's context. We proposed and evaluated two methods to find the conjunction: syntactic comparison that was trying to find conjunction between the query's keywords and each application's context keywords and semantic comparison that was finding conjunction between the query's keywords' semantically linked words and each application's context. Semantically linked words were just synonyms and hypernyms of actual query.

Finally, we select few the most relevant keywords that should expand query. To find out relevancy of each keyword we proposed two weighting methods: weight related to the application that should reflect how much is the keyword important for the current activity context, and weight related to the query that should reflect how much is the keyword related to the actual query.

3. Results

To validate our methods we proposed an experiment. 6 technical university students were told to log their activity while doing theirs normal tasks for 6 weeks. We asked them also to click on the application that is related to the actual search. Firstly, we evaluated our hypothesis that an application that is related to a query does exist. Users explicitly selected that an application is related to 88% queries so we can assume that search need comes in most cases from an application used recently. Correctness of this hypothesis allowed us to validate our second hypothesis: search activity context is the same as, or similar to, activity context of an application the searcher has used recently. Using syntactic comparison we could find the similarity between a query and the related application for 58% queries. Using semantic comparison we were able to achieve a bit higher success-rate and find a similarity for 62% queries. Moreover, we were able to find synonyms and hypernyms for 27% queries in the related application's activity context so we can make query less ambiguous immediately by extending query with these semantically linked words.

Acknowledgement: This work is partially supported by the Slovak Research and Development Agency under the contract No. APVV- 0208-10 and by the Scientific Grant Agency of the Slovak Republic, grant No. VG1/0971/11.

References

to other papers publishing the results that are summarized here

[1] Vnenk, L., Bielikova, M.: Searcher's Activity in Standalone and Web Applications as a Source for Search Query Expansion. In Proce. of ENIC 2014, European Network Intelligence Conference, IEEE CS (2014), pp.75-80.

[2] Vnenk, L.: Activity Context-aware Personalized Search. In Proc. of 10th Student Research Conference, IIT.SRC 2014, Bratislava, STU Press (2014), pp. 131-136.

Other references

[3] Aula, A., Khan, R.M., Guan, Z..: How does search behavior change as search becomes more difficult? In Proc. of the 28th Int. Conf. on Human factors in comp. systems (CHI 10), ACM (2010), pp. 35-44.

[4] Bateman, S., Teevan, J., White, R.W.: The search dashboard. In Proc. of the ACM Annual Conf. on Human Factors in Computing Systems (CHI 12), ACM (2012), pp. 1785-1794.

[5] Holub, M., Móro, R., Ševcech, J., Lipták, M., Bieliková, M.: Annota: Towards enriching scientific publications with semantics and user annotations. *D-Lib Magazine*, (2014), to appear.

[6] Kamvar, M., Kellar, M., Patel, R., Xu, Y.: Computers and iphones and mobile phones, oh my!, In Proc. 18th Int. Conf. World Wide Web (WWW 09), ACM (2009), pp. 801-810.