A User-Centered Approach for Evaluating Query Expansion Methods

Jean-Yves Delort
LIRMM – Montpellier 2 University – France
delort@lirmm.fr

Abstract

Search engines are powerful tools to find information on the Web. However, they commonly return a lot of irrelevant documents when the users' queries are not specific enough. To refine the scope of their searches, refinement terms are sometimes recommended to the users by query expansion systems (QES). A recent wide-scale survey has shown that users seldom include these terms in their queries. In this article, we propose a new user-centric approach for evaluating QES. The purpose of our approach is to assess the suggestive power of the suggested terms. Several existing QES are compared with respect to the proposed criteria.

1. Introduction

Web search engines provide their users with an increasing number of tools that assist them to reformulate their queries. Query expansion, spelling correction, "search for similar result", or "related queries" are among commonly proposed systems. Query-expansion (or query-refinement) consists in narrowing the scope of a search by adding new terms to a user's initial query.

The actual interest of the users for query expansion tools is controversial: In spite of a proven potential effectiveness in laboratory, recommended refinement terms are often ignored when tested on a wide-scale [1]. This article studies the causes of these disappointing results and suggests new evaluation criteria of query expansion systems (QES).

Relevance is the main criterion to evaluate the effectiveness of QES. It is usually characterized by the recall and precision values. This approach is often criticized because it oversimplifies the concept of relevance [2,3]. Furthermore, the actual interest of refinement terms for end-users may depend on other criteria such as:

1. the *clarity*, i.e. the fact that the word remains understandable without context,
2. the *discrimination power*, i.e. the ability to decrease the size of the next result set,
3. the *suggestive power*, i.e. the ability to help the user to improve her conceptual model or to suggest better or alternative search strategies.

In this article we assume that the content of the documents accessed during the search process often contains a couple of highly suggestive words that the user wants to refine her next query with[1]. Thus, a QES should help the user to locate these words or it should recommend them to her.

The article is organized as follow: In the next section we survey existing approaches for the evaluation of QES. In section three we present a new approach for evaluating recommended terms. The sequel of the paper compares several existing query expansion methods with this approach. Section four describes a survey of Web search behavior based on real-users. Section five introduces the word-weighting functions and presents their results.

2. Related Work

In this section we survey approaches for evaluating query expansion systems (QES).
Relevance is the main criterion to evaluate the effectiveness of QES. The usual approach consists in comparing the recall and precision values of the results retrieved for original queries with the recall and precision values of the results retrieved for the same queries enhanced with suggested terms [4]. Another approach is described in, for example [5], to evaluate methods that generate recommendations of related queries from a query log. The evaluations are based on the judgments of experts assessing the relevance of the "related queries" to the original queries. However, relevance is a multi-facetted concept. Indeed, two types of relevance can be distinguished [2]: The 'document topicality' represents the relation of a document to the topic of the user's need and the 'psychological relevance' that is "subjective and conditioned by the user's context and experience at a particular time". Recall and precision only measure the document topicality. Furthermore the recall and precision measures depend on the experts' subjectivity [3].

[1] [1] We will use the feminine pronoun (she, or) when referring to users of either gender.

End-user-based evaluations usually involve more participants than expert-based evaluations. Thus, their results tend to be more objective and accurate. On the other hand, end-user-based evaluations are based on the user explicit feedbacks [6]. Accordingly, they suffer from the fact that Web users are often reluctant to send their feedbacks, because of privacy concerns or because it is time-consuming. Furthermore, the users that agree to participate to Web surveys may represent a certain kind of web searchers.

[1] proposes an extensive evaluation of a QES based on the users' implicit feedbacks. The effectiveness is characterized by two following numbers: the number of queries followed by a user access to a result page and the number of *refined* queries followed by a user access. However, [1] notes that the results may be biased by the system interface.

In the next section we propose a new evaluation approach based on the implicit feedbacks of end-users. In this approach, the user needs not to know the recommended terms.

3. Methodology

Existing research on information-seeking behavior points out the influence of the intermediate accessed documents over the user's needs and strategies during the search process [7]. A consequence is that refined queries frequently contain terms occurring in previous pages. These new terms have a suggestive power on the user's cognitive state. We define a *borrowing query* as a query containing a new term that occurred in a previously accessed page. A query expansion system (QES) can be used to predict the refinement words added to the user's borrowing queries. In our approach we assume that the number of accurate predictions provides an estimation of the ability of a QES to recommend terms that are suggestive.

The interest in predicting terms used in borrowing queries depends on the effectiveness of these queries. Similarly to [1], we consider that a user click on a result can be interpreted as an implicit indicator of the effectiveness of the query for the user. In the sequel, we say that a query is *followed* if the user has clicked on a link of the result list. Thus, the average number of *borrowing and followed queries* and the average number of followed queries have to be compared in our approach.

For example, in Figure 1, query 3 is borrowed; queries 2 and 3 are followed. Query 1 is not followed.

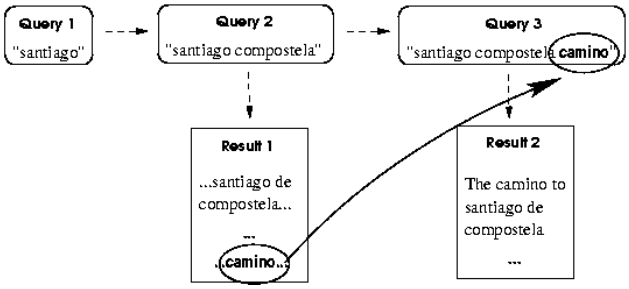

Fig 1. Borrowing and followed queries.

4. Experimentation

To test the suggestive power of query expansion algorithms, we designed a study in which users' queries to Google and users' access to result pages where recorded. The 60 participants of the survey were all PhD students and researchers in computer science. The user data were collected during 3 months.

Data were gathered with a proxy server and a redirection script. Subjects set up a proxy configuration script on their browsers so that each request to a Google page was redirected to the proxy server. The proxy server, Privoxy[2], tracks the user's IP, the requested URL, the user query (which is extracted from the URL), the referrer page and the current date and time.

It is possible to tune Privoxy to rewrite the URL contained in the pages going through it. For example, is a listing page of Google points to the link "http://www.lirmm.fr", then Privoxy can prefix it with "http://webia.lip6.fr/~delort/cgi-bin/redirect.pl?URL=". Accordingly, when the user clicks on this link, the redirection script sends a HTTP 302 REDIRECTION page to http://www.lirmm.fr to the browser. In the same time, the script saves the user's IP, the requested URL, the referrer, and the current date and time.

The contents of web pages frequently change so a daemon frequently downloads the pages of the accessed URLs to keep the real accessed content. To rebuild the user sessions, the user queries and accessed result pages are extracted from the collected data using the IP address. Parallel searches are untangled using the referrer.

[2] http://www.prixovy.org

5. Evaluation

This section presents a comparison of existing word-weighting functions with respect to the evaluation criteria introduced in section 3.

5.1. Word-weighting functions

Expansion terms recommended by query expansion systems rank expansion often use word-weighting functions. Examples of common word-weighting functions are *F4, F4mod, porter, emim, wpq,* and *zoom* [6]. These functions are based on four variables defined with respect to a corpus of documents D and a sample of documents $E \subset D$ that the user has evaluated relevant to her information need. The first two variables are N and R that represent the sizes of D and E respectively. The last two variables are defined with respect to a candidate term t: n (resp. r) is the number of documents of D (resp. E) containing t. For example the *porter* word-weighting function is:

$$w_t = \frac{r}{R} - \frac{n}{N}$$

Smeaton's query expansion system [8] is based a modification of the *porter* word-weighting function:

$$w_t = \frac{r}{R}$$

In order to evaluate n and N, we used 100000 pages belonging to the *.edu* domain in the sample of Web pages provided for the Google Programming Contest[3].

We evaluated the six word-weighting functions described in [6] and Smeaton's. Each method is given the K last accessed results as the set of candidate terms.

The function implemented in Conqueries was also tested [9]:

$$w_t = F(t) \times Q(t)$$

where $F(t)$ denotes the number of documents among the K last accessed documents that contain t and $Q(t)$ denotes the number occurrences of term t in the K last accessed documents.

5.2. Results

The log file contains 7000 user queries and user accesses to result pages made from 57 different IP addresses. The average number of terms by query is 2.63 which is close to the number reported in [10]. In contrast, the average number of accessed result pages (71.28%) is lower than the figure found in [10] (78%). The difference may be accounted for by the fact that participants make a sharper analysis of the result titles and snippets contained in the lists. It is a typical behaviour of a population used to Web search. Once Javascript and HTML has been removed, the average number of terms in the accessed results is 220.

Borrowing queries are detected using the K last accessed results. Then, the word weighting algorithms are given the K last accessed results preceding the borrowing queries. The top-M terms having the best weights are compared to the expansion terms of the borrowing queries.

Due to space limitation, this paper only presents the prediction rates of Conqueries's, Smeaton's and the best one of [6], i.e. *porter*. Figure 2 summarizes the prediction rates with respect to K and M. The X coordinate corresponds to M and the Y coordinate corresponds to the percentage of accurate prediction.

Conqueries's function has a percentage of accurate prediction slightly better than other methods. When 10 terms can be recommended (i.e. M=10), and the two last accessed results are used (K=2), the probability that a term is borrowed (if there is one), is 4.8. It can be compared to a random selection: 10 * 100/440=2.2 (if one assumes that there is only one term borrowed out of 220*2=440).

The experimentation also brought out the two following results:

- When K=2, 15.7% of queries are borrowing.
- 35.55% of the queries are followed, but this percentage rises to 50% when the queries are borrowing terms picked up among the K=5 last results.

The first result suggests that a really good query expansion system is able to make up to 15.7% of relevant predictions with a data source containing the two last accessed results. The second result implies that such predictions tend to be really effective from the user point of view.

[3] http://catalogs.google.com/programming-contest/

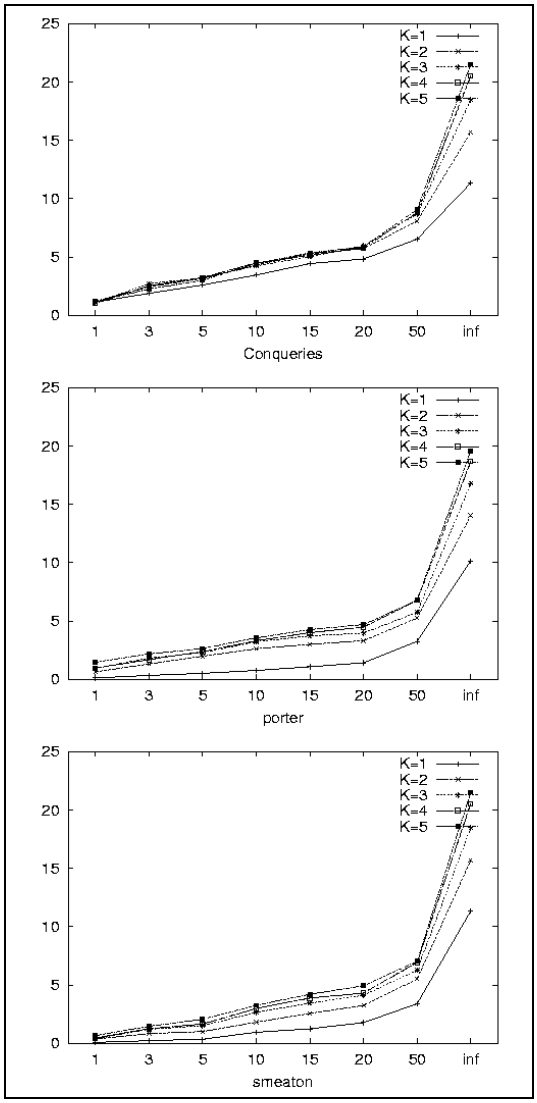

Fig 2. Prediction rates

6. Conclusion and Future Work

Refining the queries submitted to search engines is an efficient strategy to get ride of the irrelevant pages contained in the result lists. The actual effectiveness of current query expansion systems (QES) is controversial. In this paper we have suggested several reasons accounting for it: Firstly, current approaches aim at improving the relevance of the result sets but they use a limited representation of relevance. Secondly, they do not take into account the clarity, or the discriminating power or the suggestive power of the terms.

The paper describes a new approach to evaluate the effectiveness of QES. The proposed approach aims at assessing the suggestive power of the terms recommended by a QES. The user's querying behavior plays a central role in the evaluation. A survey was conducted with real-users to compare several existing QES with respect to this approach. The analysis of the results highlights two results of general interests: 1) many queries contain terms picked up in the content of the previous accessed contents and 2) such queries are judged more effective from the user's point of view.

We are currently working on a new survey intended to compare the actual effectiveness of the QES with the effectiveness assessed by our approach using explicit user feedback.

7. References

[1] P. Anick, "Using terminological feedback for web search refinement - a log-based study", *Proceedings of the Conference on Research and Development in Information Retrieval*, 2003, pp. 88-95.

[2] M.H. Heine, "Reassessing and Extending the Precision and Recall Concepts", *Proceedings of MIRA'99*, 1999.

[3] T. Saracevik, "Information Science", 1999, *Journal of the American Society of Information Science*, 50 (12), 1999, pp. 1051-1063.

[4] C.J. Van Rijsbergen, *Information Retrieval*, London: Butterworths, 1979

[5] B.M. Fonseca et al., "Using Association Rules to Discover Search Engines Related Queries", *Proceedings of The First Latin American Web Congress*, 2003, pp. 66-71.

[6] E.N. Efthimiadis, "A user-centred evaluation of ranking algorithms for interactive query expansion", Proceedings of the 16th *Conference on Research and development in information retrieval*, 1999, pp. 146-159.

[7] M.J. Bates, "The design of browsing and berrypicking techniques for the online search interface", *Online Review*, 13(5), 1989, pp. 407-431.

[8] A.F. Smeaton and F. Crimmins, "Relevance feedback and query expansion for searching the web: A model for searching a digital library", *Proceedings of The European Conference on Digital Librairies*, 1997, pp. 99-112.

[9] J.Y. Delort, "CONQUERIES: An Agent that Assists Query Expansion", *Proceedings of the Fifth International Conference on Web Engineering*, 2005.

[10] B.J. Jansen, A. Spink, J. Bateman and T. Saracevic, "Real Life Information Retrieval: A Study of User Queries on the Web", *SIGIR Forum*, 32 (1), 1998, pp. 5-17.