QUINN: Query Updates for News Monitoring

Suzan Verberne
Radboud University, The Netherlands

Thymen Wabeke
TNO, The Hague, The Netherlands

Rianne Kaptein
TNO, The Hague, The Netherlands

LexisNexis Publisher[1] is an online tool for news monitoring. Organizations use the tool to collect news articles relevant to their work. For monitoring the news for a user-defined topic, LexisNexis Publisher takes a Boolean query as input, together with a news collection and a date range. The output is a set of documents from the collection that match the query and the date range.

For the users it is important that no relevant news stories are missed. Therefore, the query needs to be adapted when there are changes to the topic. This can happen when new terminology becomes relevant for the topic (e.g. 'wolf' for the topic 'biodiversity'), there is a new stakeholder (e.g. the name of the new minister of economic affairs for the topic 'industry and ICT') or new geographical names are relevant to the topic. (e.g. 'Heumensoord' for the topic 'refugees') The goal of the current work is to support users of news monitoring applications by providing them with suggestions for query modifications in order to retrieve more relevant news articles.

The user can control the precision of the final publication list by disregarding irrelevant documents in the selection. Recall is more difficult to control because the user does not know what he has not found. Our intuition is that documents that are relevant but *not* retrieved with the current query have similarities with the documents that *are* retrieved by the current query. Therefore, our approach to query suggestion is to generate candidate query terms from the set of retrieved documents. This approach is related to pseudo-relevance feedback, a method for query expansion that assumes that the top-k retrieved documents are relevant, extracting terms from those documents and adding them to the query. There are three key differences with our approach: First, instead of adding terms blindly, we provide the user with suggestions for query adaptation. Second, we take into account an important characteristic of news data: the collection is constantly changing. We hypothesize that terms that show a big increase in frequency over time are candidate new query terms, because they were not relevant in an earlier stage of the news stream. Third, we have to deal with Boolean queries, which implies that we do not have a relevance ranking of documents to extract terms from. This means that the premise of 'pseudo-relevance' may be weak for the set of retrieved documents.

Our approach for query term extraction is as follows: For a given Boolean query, we retrieve the result set R_{recent}, which is the set of articles published in the last 30 days, and the result set R_{older}, which is the set of articles published 60 to 30 days ago. We implemented four different term scoring algorithms from the literature, and used each of them to extract three term lists: T_1 is the divergence between R_{recent} and a generic news background corpus; T_2 is the divergence between R_{recent} and R_{older}; T_3 is the divergence between R_{older} and the generic news background corpus. The query suggester returns one of three term lists to the user: $A = T_1$; $B = T_2$ and $C = \{t : t \in T_1 \wedge t \notin T_3\}$.

The demo application has been used to collect feedback from expert users of LexisNexis Publisher to determine the best method for generating term suggestions. In the application, a Boolean query can be entered that is used to search in Dutch newspapers. The found documents are shown in a result list and a list of query term suggestions (a pool of terms from all methods) is presented. Users were asked to judge the relevance of the returned terms on a 5-point scale, could update the search query (potentially with a suggested term) and retrieve a new result list.

The results of our user experiment show that with the best performing method (method A with either Parsimonious Language Models or Kullback-Leibler Divergence as term scoring algorithm), the user selected a term from the top-5 suggestion list for only 13% of the topics, and judged at least one term as relevant (relevance score >= 4) for 25% of the topics. Inspection of the results and the user comments revealed that the term suggestions are noisy, mainly because the set of retrieved documents for the Boolean query is noisy. We expect that the use of relevance ranking instead of Boolean retrieval, and a post-filtering for noisy terms, will give better user satisfaction.

Acknowledgements

This publication was supported by the Dutch national program COMMIT (project P7 SWELL).

[1] http://www.lexisnexis.com/bis-user-information/publisher/