TopicForest: A Prototype Discovery Engine

Soheil Danesh (soheildb@gmail.com)

I. Introduction

It is generally accepted that there are three types of web search queries: Navigational, transactional and informational [1]. Navigational queries occur when users want to navigate to specific web destinations, for example their favorite news websites. Transactional queries describe users who intend to make a web-mediated transaction and are looking for a few trustworthy parties for it. Searching for restaurants nearby could be seen as an example of a transactional query. Informational queries are issued by users who want to learn about a new, potentially broad topic such as 'Cancer' or 'Middle East Politics'. In this case, due to the wide scope of the queries, the desired information might not reside in a single or few web documents but rather be scattered across many documents. In fact there could be several subtopics or aspects in the query domain [2] which might not be discussed in the same documents.

Traditional Search Engines rank documents containing the query terms by relevance and are most effective when the few top ranked documents contain the required information. This model generally works very well for navigational and transactional queries by providing the users with the single or few desired web documents. However it does not fully meet the needs of informational queries because, as discussed above, the results of such broad queries tend not to be contained in a single or few documents. More importantly, the domain of an informational query may have different subtopics, some of which might not be covered at all in the top ranked search results of traditional Search Engines.

In the case of broad informational queries, one method to address such shortcomings of traditional Search Engines could be through a new paradigm of presenting search results: Instead of returning a ranked list of documents, a hierarchical topical summary of the document set could be returned. For a given query the user can interactively navigate the topic hierarchy and gain an overview of the different topics in the domain of query before narrowing down on those of interest and eventually reaching individual documents. This provides a holistic view of the broad query domain and decreases the likelihood of users missing important sub-domains of the query subject. Such a tool could be called a Discovery Engine as opposed to a Search Engine. A Search Engine is most effective at returning one or few documents containing the specific information the user is searching for. On the other hand, the Discovery Engine is designed to help users discover interesting information about broad topics in large document sets.

II. Method and Evaluation

To experiment with the new paradigm of presenting search results as hierarchical topical summaries a prototype Discovery Engine has been developed and is available online at http://topicforest.com. TopicForest presents the results of PubMed searches as topical hierarchies called "Topic Trees". Users can interactively explore the search results through the topic tree by clicking on each key term (I.e. Topic). Doing so presents the user with the set of documents containing that term and all of its parent topics in the tree. This allows the user to gain a higher level view of the topics involved in the query domain before traversing down a tree branch to zoom in on increasingly specific topics. The terminology extraction algorithm used to produce the topic hierarchies is based on as of yet unpublished improvements made to unsupervised keyphrase extraction methods previously presented in [3]. This algorithm is fully unsupervised and is used with its default parameters in the following evaluation. To evaluate the topically summarized search results the query "Covid-19" was submitted to TopicForest on May 22. (Results: http://topicforest.com/queryterm/covid$_1$9/$topictree$). In this evaluation we stick to only the first level topics of the topic tree. To evaluate the validity of the topics a medical doctor was asked to manually label each topic as relevant or irrelevant to Covid-19. Out of 29 first level topics only 4 were deemed not directly relevant to Covid-19. These first level terms can be seen below, sorted in the order returned by TopicForest. Terms deemed irrelevant have been underlined.

Severe acute respiratory syndrome coronavirus, Personal protective equipment, Critically ill patients, Social media, Urgent need, Corona virus disease, Mortality rate, Study aimed, Logistic regression, **CT findings, Transplant recipients,** Confidence interval, **Psychological distress, Since December, New York City, Symptom onset, Social distancing measures, Converting enzyme, Ground glass, Physical distancing, Common symptoms, Provide guidance, Early stages, South Korea**, Literature search, **Nuclear medicine, Retrospective cohort, Radiation oncology, General population**

As a comparison we manually assigned topics to the top 20 ranked results returned by PubMed for the same query. We discovered several of the above topics such as personal protective equipment, Social Media, Psychological distress, transplant recipients, Nuclear Medicine and Radiation Oncology among others were not represented at all. Topics found only in the search results and not the TopicForest were "Dental care" and "Virus Reproduction Number".

REFERENCES

[1] A. Broder, "A taxonomy of web search," in *ACM Sigir forum*, vol. 36, pp. 3–10, ACM New York, NY, USA, 2002.

[2] M. J. Welch, J. Cho, and C. Olston, "Search result diversity for informational queries," in *Proceedings of the 20th international conference on World wide web*, pp. 237–246, 2011.

[3] S. Danesh, T. Sumner, and J. H. Martin, "Sgrank: Combining statistical and graphical methods to improve the state of the art in unsupervised keyphrase extraction," in *Proceedings of the fourth joint conference on lexical and computational semantics*, pp. 117–126, 2015.