Exploiting Temporal Information in Retrieval of Archived Documents

Nattiya Kanhabua
Dept. of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway
nattiya@idi.ntnu.no

ABSTRACT

In a text retrieval community, many researchers have shown a good quality of searching a current snapshot of the Web. However, only a small number have demonstrated a good quality of searching a long-term archival domain, where documents are preserved for a long time, i.e., ten years or more. In such a domain, a search application is not only applicable for archivists or historians, but also in a context of national library and enterprise search (searching document repositories, emails, etc.). In the rest of this paper, we will explain three problems of searching document archives and propose possible approaches to solve these problems. Our main research question is: How to improve the quality of search in a document archive using temporal information?

One problem of search is a large amount of documents retrieved and this results in the decreasing accuracy since users have to spend more time in exploring the information needed. Possible solutions can be re-ranking for presentation or including a temporal relationship with respect to a query, i.e., extending keyword search with a creation or update date of documents (called temporal criteria). In that way, a system narrows down a set of results by retrieving documents according to both text and temporal criteria, e.g., temporal text-containment search [1]. Two ways to obtain temporal criteria relevant to a query are 1) having them provided by users, or 2) determined by the system. The former way is already done in some works. So, we propose to attach a time or time period to the query implicitly, and retrieve (or give preference to) results created within that time(-period). Note that, sometimes users have no clue regarding possible time of documents, thus no time can be explicitly presented in the query. Besides, this technique is only useful for some specific queries where key words are *strongly time-related*, e.g., by looking at words statistics, etc. We can determine time of queries based on word usages by using our temporal language model based time-determination [2].

Another problem is the effect of language changes over time, e.g., changes of words related to their definition, semantics, and name (people, location, etc.). In some cases, original words are obsolete, for example, before the year 1939, the name Siam was used for Thailand and it is rarely used nowadays. This makes trouble if both query and documents are represented in different forms, i.e., historical or modern forms. We propose to handle this problem during query time by using a dictionary linking concepts and entities based on time. Thus, for a query for "Thailand", the query might be expanded to "Thailand or Siam". For a query for documents written at a certain time (before 1939), the query might be rewritten from "Thailand" to "Siam". The expansion has been done before in two ways: an expansion of query and an expansion of index. In the first case, they automatically construct rules for mapping historic terms into modern terms. In the latter case, based on a lexical database, they index synonyms and holonyms as additional indices of a term. We propose to explicitly make use of time in a term mapping process in order to improve the search quality concerning the language changes. This can achieve by building a time-concept dictionary from the well-known and freely available encyclopedia, i.e., Wikipedia.

In general, when searching in news, hit-list documents are displayed in a chronological order where newer pages are more important/relevant than older ones. However, a chronological filtering is not always needed. Therefore, the ordering of documents by taking into account temporal information, i.e., temporal ranking is necessary. We propose to analyze a document collection to obtain a topical trend (the trend of a topic) that can be represented as the weight of a topic over time. For example, a document about "tsunami" written in 2004 should receive a higher weight than that written in 2008. For a given query, documents will be retrieved based on their similarity scores, e.g., TF-IDF to a query topic. However, the ranking of documents is the combination of their similarity scores and document weights with respect to a topical trend. In fact, the two proposed approaches above retrieve documents relevant to temporal criteria of a query, called *a query temporal profile*. On the other hand, the temporal ranking approach retrieves documents relevant to a query topic and ranks them based on their *document temporal profiles*.

Categories and Subject Descriptors: H.3.3[Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Experimentation, Measurement

Keywords: Language Models, Query Expansion, Ranking, Temporal Search

1. REFERENCES

[1] K. Nørvåg. Supporting temporal text-containment queries in temporal document databases. *Journal of Data & Knowledge Engineering*, 49(1):105–125, 2004.

[2] N. Kanhabua and K. Nørvåg. Improving Temporal Language Models for Determining Time of Non-timestamped Documents. In *Proceedings of ECDL'08*, 2008.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.