DOCTORAL ABSTRACT

Disambiguating Implicit Temporal Queries for Temporal Information Retrieval Applications

Ricardo Campos
Polytechnic Institute of Tomar
Portugal
LIAAD–INESC TEC
rncampos@inescporto.pt

Time is an important dimension of the information retrieval area that can be very useful in helping to meet the users information needs whenever they include temporal intents. However retrieving the information that meets the query demands is not an easy process. The ambiguity of the query is traditionally one of the causes impeding the retrieval of relevant information. This is particularly evident in the case of temporal queries where users tend to be subjective when expressing their intents (e.g., avatar movie instead of avatar movie 2009). Determining the possible times of the query is therefore of the utmost importance when attempting to achieve better disambiguated results and in order to enable new forms of exploring them.

In this thesis, we present our contributions to disambiguate implicit temporal queries in real-world environment, i.e. the Web. To understand better this type of queries, three directions may be followed: information extracted from (1) metadata, (2) query logs or (3) document contents. Within the context of this thesis, we will focus on the latter. However, unlike existing approaches we do not resort to a classification methodology. Instead, in our approach, we seek to detect relevant temporal expressions based on corpus statistics and a general similarity measure that makes use of co-occurrences of words and years extracted from the contents of the documents. Moreover, our methodology tends to be mostly language-independent as we do not use any linguistic-based techniques. Instead, we use a rule-based model solution supported by regular expressions.

Based on this, we start by performing a comprehensive study of the temporal value of web documents, particularly web snippets, showing that this type of collection is a valuable data source in the process of dating implicit temporal queries. We then develop two methods. A temporal similarity measure to evaluate the correlation between the query and the candidate dates identified, called Generic Temporal Evaluation (GTE) and a threshold-based classifier that selects the most relevant dates while filtering out the non-relevant or incorrect ones, known as GTE-Class. Subsequently, we propose two different applications named GTE-Cluster and GTE-Rank. The first one, uses the determined time of the queries to improve search results exploration. For this purpose, we propose a flat temporal clustering model solution where documents are grouped at the year level. GTE-Rank, in turn, uses the same information to temporally re-rank the web search results. We employ a combination approach that considers words and temporal scores, where documents are ranked to reflect the relevance of the snippet for the query, both in the conceptual and in the temporal dimension.

Through extensive experimental evaluation, we mean to demonstrate that our models offer promising results in the field of Temporal Information Retrieval (T-IR), as demonstrated by the experiments conducted over web corpora. As an additional contribution to the research community, we publicly provide a number of web services so that each of the different approaches can be tested. Although the main motivation of our work is focused on queries with temporal nature, the implemented prototypes allow the execution of any query including non-temporal ones. Finally, for future research direction, we study the behavior of web snippets in the context of Future Information Retrieval (F-IR), a fairly recent topic which consists of extracting future temporal information in order to answer user queries with a future temporal nature.

Available online at `http://www.ccc.ipt.pt/~ricardo/PhDThesis_RCampos.pdf`