Information Processing and Management 52 (2016) 273–298

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

GTE-Rank: A time-aware search engine to answer time-sensitive queries

Ricardo Campos[a,b,*], Gaël Dias[d], Alípio Jorge[b,c], Célia Nunes[e,f]

[a] Polytechnic Institute of Tomar, Tomar, Portugal
[b] LIAAD/INESC TEC – INESC Technology and Science, Portugal
[c] DCC – FCUP, University of Porto, Portugal
[d] HULTECH/GREYC, University of Caen Basse-Normandie, Caen, France
[e] Department of Mathematics, University of Beira Interior, Covilhã, Portugal
[f] Center of Mathematics, University of Beira Interior, Covilhã, Portugal

ARTICLE INFO

Article history:
Received 19 August 2014
Revised 23 July 2015
Accepted 31 July 2015
Available online 10 November 2015

Keywords:
Temporal information retrieval
Time-sensitive queries
Temporal re-ranking
Temporal query understanding

ABSTRACT

In the web environment, most of the queries issued by users are implicit by nature. Inferring the different temporal intents of this type of query enhances the overall temporal part of the web search results. Previous works tackling this problem usually focused on news queries, where the retrieval of the most recent results related to the query are usually sufficient to meet the user's information needs. However, few works have studied the importance of time in queries such as "Philip Seymour Hoffman" where the results may require no recency at all. In this work, we focus on this type of queries named "time-sensitive queries" where the results are preferably from a diversified time span, not necessarily the most recent one. Unlike related work, we follow a content-based approach to identify the most important time periods of the query and integrate time into a re-ranking model to boost the retrieval of documents whose contents match the query time period. For that purpose, we define a linear combination of topical and temporal scores, which reflects the relevance of any web document both in the topical and temporal dimensions, thus contributing to improve the effectiveness of the ranked results across different types of queries. Our approach relies on a novel temporal similarity measure that is capable of determining the most important dates for a query, while filtering out the non-relevant ones. Through extensive experimental evaluation over web corpora, we show that our model offers promising results compared to baseline approaches. As a result of our investigation, we publicly provide a set of web services and a web search interface so that the system can be graphically explored by the research community.

1. Introduction

Despite the growing importance of time in information retrieval, most of the existing ranking functions are limited to simply returning the freshest results (Berberich, Vazirgiannis, & Weikum, 2005; Cheng, Arvanitis, & Hristidis, 2013; Dai, Shokouhi, & Davison, 2011; Dong et al., 2010; Efron & Golovchinsky, 2011; Li & Croft, 2003; Zhang, Chang, Zheng, Metzler, & Nie, 2009). Current search engines for example, either give users the possibility to specify a point-in-time of their interest or apply freshness metrics to push to the top list the most recent results. While this may be a suitable solution for the news domain for which a huge

[*] Corresponding author at: Polytechnic Institute of Tomar, Tomar, Portugal.
E-mail addresses: ricardo.campos@ipt.pt (R. Campos), gael.dias@unicaen.fr (G. Dias), amjorge@fc.up.pt (A. Jorge), celian@ubi.pt (C. Nunes).

http://dx.doi.org/10.1016/j.ipm.2015.07.006
0306-4573/© 2015 Elsevier Ltd. All rights reserved.

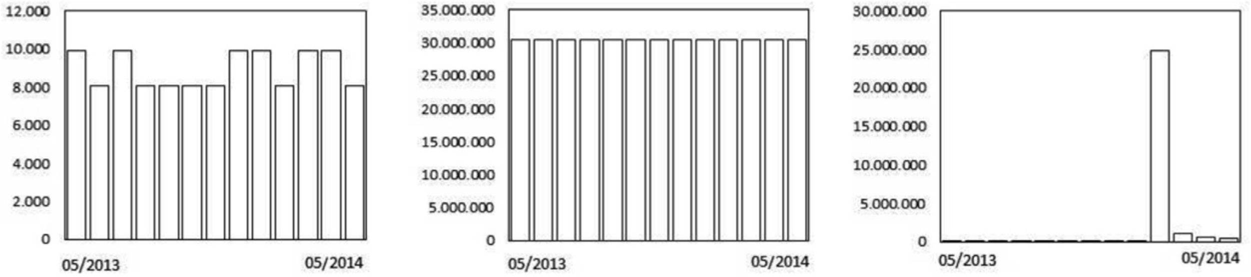

Fig. 1. Average number of monthly queries over the period 06/2012–05/2014 for (a) “first moon landing”. (b) “WWW”. (c) “Philip Seymour Hoffman”. Adapted from Google AdWords.

quality of time-stamped web pages are available and for recent events which require evidence spike phenomena, it may prove to be inefficient if the user is more interested in information covering a broader timespan. For instance, a user specifying the query “Philip Seymour Hoffman” on February 2014 is likely to be interested in web pages related to the death of this well-known American actor, yet the user might be also looking for Hoffman’s biography for which non-fresh documents are sufficient. The same query issued a few months later, will probably be better answered with information from different time periods, such as: when was the actor born, or when did be begin acting in television. In this case a wide coverage of the information required will be more appropriate, yet current approaches still favor more recent documents.

Aware of this, researchers have started to address the problem of returning documents that are not only topically relevant but that are also from the most important time periods and not just the latest. In order to tackle this problem a few works have been introduced. The methods proposed to solve this problem can be broadly divided into two classes: (i) metadata-based approaches and (ii) query-log based approaches. One family of methods exploits the publication date of the document to identify the most important time periods of the query thereafter using this information to promote results around that time frame. Other approaches rely on volume-based techniques or similar related user queries (e.g. “Philip Seymour Hoffman 2014”) to favor documents matching the determined time of the query. For a large number of scenarios however, this is no solution. Firstly, the timestamp of a document (creation, publication, or modification time) may differ significantly from its focus time, i.e., its content. A simple example is a document published in “*2009*” whose content concerns the year “*2011*”. In addition, metadata information is particularly difficult to obtain from less structured collections, such as web pages, as opposed to news articles. One reason for this, as observed by Nunes, Ribeiro, and David (2007), is due to the fact that web servers typically do not provide other temporal information than the crawling date.

Secondly, although relying on web query logs may be a straightforward solution to infer the temporal value of time-sensitive queries, access to real-world query logs outside large industrial labs is difficult and a huge impediment to information retrieval research (Callan & Moffat, 2012). A further challenge is that extracting temporal information from web query logs implies one of two things: (i) the previous issue of similar related queries or (ii) the occurrence of spikes in the number of queries issued. In the first case we face a query-dependency problem compounded by the fact that only around 1.2% of the queries are temporally explicit by nature (Campos, Dias, & Jorge, 2011). This constitutes a handicap to infer the time frame of a time-sensitive query. Moreover, as stated by Campos et al. (2011), the mere fact that a query is year-qualified does not necessarily mean that it has a temporal intent (e.g., “*Microsoft office 2007*”) or that the associated year is actually correlated with the query (e.g., “*football World Cup 2012*” – there was no World Cup in 2012). As an alternative solution to using similar queries, volume-based techniques can be applied acting as a clue of the queries’ timeliness. However, these techniques are dependent on the query volume and on the distribution of queries over time. For a large number of queries this solution is simply unfeasible. In particular, several queries may not exhibit any spike, remain steady over time or may not necessarily reflect the different temporal dimensions of the query. Indeed, the number of queries issued throughout time is highly correlated with the users’ demands, which might negatively affect a clear understanding of the entire picture. A representation of this is given in Fig. 1 for the query “first moon landing”, “WWW” and “Philip Seymour Hoffman”, where vertical bars represent the average number of monthly queries issued on Google commercial search engine for the period 2013–2014. A quick look at the figure shows that, for different reasons, user searches are not sufficient to help in understanding the different time periods of the queries. “WWW” (see Fig. 1b) portrays the example of queries for which user searches remain steady over time. “first moon landing” (see Fig. 1a) and “Philip Seymour Hoffman” (see Fig. 1c) queries depict, in contrast, the example of cases where particular events, such as the landing happening or Hoffman’s birthdate, cannot simply be inferred from web logs, due to the inexistence of Internet records as of the date of these happenings.

To address the above shortcomings, we make use of web contents to infer the temporal nature of implicit temporal queries. That is, we identify relevant temporal expressions from web snippets related to the query, thereafter using this information to improve the quality of the results retrieved. Timeliness is then incorporated in a ranking model through a linear combination of topical and temporal scores, thereby reflecting the relevance of any web document both in the topical and temporal dimensions. The rationale is that offering the user a comprehensive temporal contextualization of the topic is intuitively more informative than simply retrieving only the most recent results or just its topical perspective. Experiments with two publicly