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.

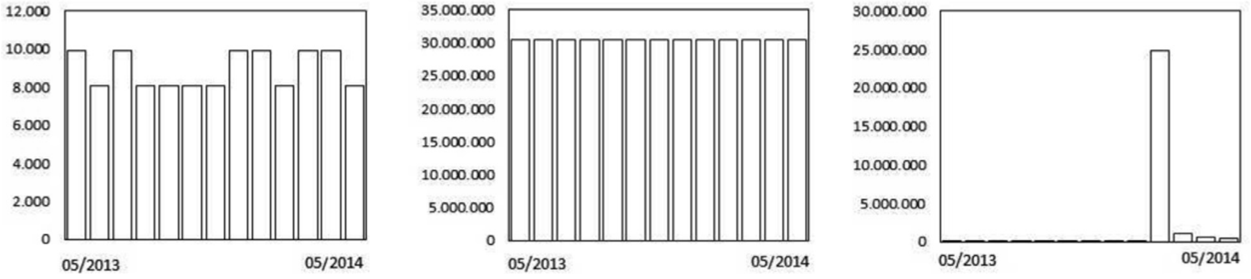


Fig. 1. Average number of monthly queries over the period 06/2012–05/2014 for (a) "first moon landing". (b) "WWW". (c) "Philip Seymour Hoffman". Adapted from Google AdWords.

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 available datasets show that the results improve when GTE-Rank (our proposed model) is applied. This can be very useful for a large set of timely underspecified queries, which although not explicitly temporally tagged, still have an inherent implicit temporal nature.

The contributions of this research can be summarized as follows: (1) we introduce a novel temporal re-ranking function based on the identification of top relevant dates for queries where no temporal criteria is provided; (2) we adopt a language-independent methodology (as long as Occidental languages are concerned) that can be applied to real-world search scenarios; (3) by using a content-based approach, we manage to return documents about a given period, as opposed to the retrieval of documents written or published at a given date; (4) we provide public access to a set of queries, web documents and ground-truth results, so that our evaluation outcomes can be compared with future approaches and (5) we divulge a few web services and a user search interface so that GTE-Rank can be tested and used by the research community.

The structure of this paper is as follows. Section 2 opens with a discussion of the relevant literature. Section 3 presents our top relevant dates query identification model, which builds the foundations for the temporal ranking approach described in Section 4. Section 5 introduces experimental setups. Section 6 discusses obtained results. Section 7 presents the results of a user study crowdsourcing experiment. Section 8 introduces the user search interface, the web services and presents a set of examples to evidence important features of our ranking model. Finally, Section 9 summarizes this research with some final remarks and suggestion of future research directions.

2. Related work

There is a broad range of works for temporal ranking. Most pioneering approaches have attempted to improve the exploration of search results by biased ranking functions, usually by favoring more recent documents matching the user's query. One of the first works attempting to solve this problem was developed by Li and Croft (2003). In it, the authors incorporate time into both query-likelihood and relevance-based language models. Documents with a more recent creation date are assigned a higher probability. Similarly, Berberich et al. (2005), Zhang et al. (2009) and Efron and Golovchinsky (2011) describe a re-ranking score so that fresh documents are ranked higher. The underlying assumption is that the user's intent is to find documents concerning the most recent years. Within the context of learning to rank, Dong et al. (2010) propose a retrieval system to answer recency breaking-news queries, where document freshness is taken into account by means of multiple temporal features, such as the timestamp or the link time. The paper published by Dai et al. (2011) propose a machine learning model that optimizes freshness and relevance, where weights depend on the query's temporal profile. More recently, Cheng et al. (2013) presented a language model that incorporates the timeliness factor in order to retrieve recent results for non-spike timely queries.

Other works focus mainly on time-sensitive queries where the results retrieved concentrate on specific time periods. Jones and Diaz (2007) use a language model solution and a collection of web news documents to model the period of time that is relevant to a query, where temporal information is extracted from the document timestamp. Queries are then classified into one of three classes depending on how documents spike in the collection over time. Broadly, queries can be classified as (i) atemporal queries, i.e., queries which are not sensitive to time (e.g., "*icecream recipes*"); (ii) temporal unambiguous queries, i.e., queries which are characterized by pointing to a concrete time period (e.g., "*first moon landing*"), and (iii) temporal ambiguous queries, i.e., queries which either refer to periodical events (occurring on a recurring basis, e.g., "*boston marathon*") or aperiodic events (occasional peaks of popularity lacking periodicity, e.g., "*haiti earthquake*").

In another line of research, Berberich, Bedathur, Alonso, and Weikum (2010) propose the integration of temporal expressions into a language model framework and the ranking of documents according to the estimated probability of generating the query, but it requires queries to contain an explicit temporal expression. Kanhabua and Nørvåg (2010) in turn, use a query's determined time to improve the re-ranking of the web page results. The idea behind their research is that documents with creation dates that closely match the query's time are more relevant in the temporal dimension and thus should be ranked higher. To achieve this goal, they proposed a mixture model to linearly combine both textual and temporal similarities. The method put forward by Styskin, Romanenko, Vorobyev, and Serdyukov (2011) relies on a recency-sensitive query classifier to apply result diversification by combining ordinary search results with fresh documents. Similar to Jones and Diaz (2007), Dakka, Gravano, and Ipeirotis (2012) resort on spikes in the number of documents matching the query over time to identify the most important time stamps relevant to the query. Document relevance is then estimated by incorporating time into language models. Kanhabua and Nørvåg (2012) in turn, propose a new approach by applying a time-sensitive ranking model based on learning to rank techniques to answer explicit temporal queries. To learn the ranking model, two classes of features are applied: temporal (document focus time and its timestamp) and entity-based (persons, locations, or organizations). The results obtained by this model outperform the proposed method of Berberich et al. (2010). Both, however, require the specification of an explicit temporal query and resort to the creation date of documents as the correct temporal signal, which is far from being credible or accessible in most cases. A more recent work of Berberich and Bedathur (2013) explores the concept of temporal diversification and proposes an approach in which search results are composed of documents that were also published at diverse times of interest to the query. Despite these proposals perform well in the specific context of news, a more general solution that addresses this problem by resorting to any type of documents, such as "regular" web pages is also needed.

Another branch of research supports on web query logs to boost the results around a given time frame. Metzler, Jones, Peng, and Zhang (2009) for instance, propose to discover important time periods from similar related queries by mining web query logs. A time-dependent ranking model that explicitly adjusts the score of a document in favor of those matching the users'

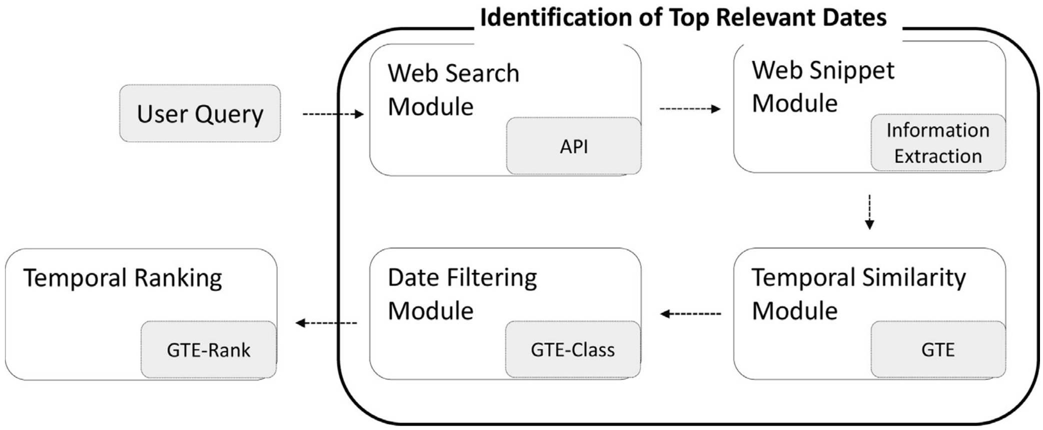


Fig. 2. Overall architecture.

implicit temporal intents is then proposed. More recently Chang, Huang, Yang, Lin, and Cheng (2012) propose to re-rank web search results making use of user temporal click information extracted from web query logs. Further other works have been proposed over the years, which are exhaustively described in Campos, Dias, Jorge, and Jatowt (2014b).

While all the above models rely on spikes in the distribution of relevant documents or queries, none extracted temporal information from web content. In this research, we propose a different strategy that follows a content-based methodology that extracts temporal features from the contents of the documents. We differ from previous studies on this subject in several other aspects. First, we do not make use of query logs. Second, we do not rely on the creation date of a document as it may differ significantly from its content. Third, our methodology is unsupervised as no specific training process is needed. Fourth, we do not resort to temporal language models, which are difficult to adapt to open domain collections due to a training process demand, thus making it possible to have an on-line solution without any restriction in terms of the query time period. Fifth, our solution is language-independent as far Occidental language is concerned as it implements a rule-based model supported by simple language-independent regular expressions to extract relevant dates from web documents. Finally, besides estimating the degree of relevance of a temporal expression, we propose to determine whether or not a date is query relevant, thus using this information to improve the re-ranking of web search results, a major difference compared to related works which consider any occurrence of temporal expressions in web documents and other web data as equally relevant to a time-sensitive query.

3. Identification of top relevant dates for time-sensitive queries

In this section, we describe the method that guides our identification of top relevant dates related to text queries with a temporal dimension. Since results are produced "on-the-fly", we simply return the set of n-top web snippets retrieved in response to the user's query, thus keeping the system computationally efficient. Indeed, it is important to notice that web snippets are an interesting alternative for the representation of web documents (Alonso, Baeza-Yates, & Gertz, 2009; Alonso, Gertz, & Baeza-Yates, 2011), where years often appear (Campos et al., 2011) thus avoiding the cost of parsing full web pages. The overall idea of the process, represented in Fig. 2, is to identify and classify years which are relevant for a given query based on four different steps which constitute the basis for the temporal ranking approach (see Section 4): (1) Web search; (2) Web snippet representation; (3) Temporal similarity and (4) Date filtering. Each one is described in the upcoming sections.

3.1. Web search

In our work, we deal with implicit temporal queries (e.g. "*football world cup*") since handling explicit temporal ones (e.g. "*football world cup 2014*") is a less complex task. We apply a web search API which, given a query q, accesses an up-to-date index search engine to obtain a collection of n web snippets $S = \{S_1, S_2, \ldots, S_n\}$.

3.2. Web snippet representation

Each S_i, for $i = 1, \ldots, n$, denotes the concatenation of two texts, i.e. $\{Title_i, Snippet_i\}$ and is represented by a bag-of-relevant-words and a set of candidate temporal expressions. In what follows, we assume that each S_i is composed by two different sets denoted W_{S_i} and D_{S_i}.

$$S_i \rightarrow (W_{S_i}, D_{S_i}), \quad (1)$$

Specifically $W_{S_i} = \{w_{1,i}, w_{2,i}, \ldots, w_{k,i}\}$ is the set of the k most relevant words/multiwords associated with a web snippet S_i, and $D_{S_i} = \{d_{1,i}, d_{2,i}, \ldots, d_{t,i}\}$ is the set of the t candidate years associated with a web snippet S_i. Moreover,

$$W_S = \cup_{i=1}^{n} W_{S_i}, \quad (2)$$

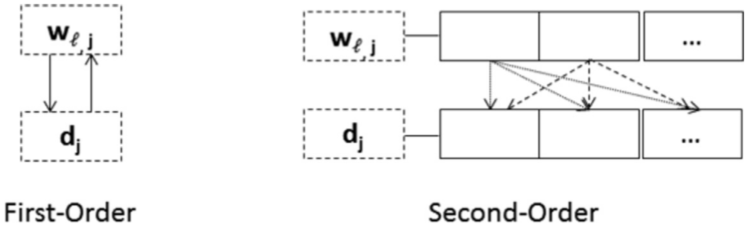


Fig. 3. Example of first-order and second-order similarity measures.

is the set of distinct relevant words/multiwords (hereafter called terms) extracted for a query q, within the set of web snippets, S i.e. the relevant vocabulary. In this research, relevant words are identified using a web service[1] provided by Machado, Barbosa, Pais, Martins, and Dias (2009), which selects terms based on a specific segmentation process and a numeric selection heuristic.

Similarly,

$$D_S = \cup_{i=1}^{n} D_{S_i}, \tag{3}$$

is defined as the set of distinct candidate years extracted from the set of all web snippets S. Given the simplicity of the task, we opt to use a self-defined rule-based model to extract the following explicit temporal patterns: *YYYY, YYYY-YYYY, YYYY/YYYY, MM/dd/YYYY, dd/MM/YYYY, MM.dd.YYYY* and *dd.MM/YYYY*. This contrasts with the use of temporal taggers (e.g., HeidelTime[2] Strötgen and Gertz (2010) and SuTime[3] Chang and Manning (2012), which are better suited to more complex tasks, yet mostly useful for only a few languages (typically English) and for one domain (usually, the news domain). For each discovered pattern, the temporal expression is then normalized to *YYYY*. Although it is possible to extract temporal expressions with finer granularities we limit our approach to the extraction of years as finding days or months without a year reference coupled together is uncommon, given the specificities of web snippets (particularly its small size collection). Note that a document can also contain other types of temporal expressions, such as implicit or relative. However, likewise months and days they seldom occur in web snippets, thus they will not be considered in this research.

Finally,

$$W_j^* = W_S \cap d_j, \tag{4}$$

is defined as the set of relevant words W_S that appear together with the candidate date d_j in any web snippet.

3.3. GTE: temporal similarity measure

We formally define the problem of (query, candidate date) temporal relevance as follows: given a query q and a candidate date $d_j \in D_S$ assign a degree of relevance to each (q, d_j) pair. To model this relevance, we will use GTE, a temporal similarity measure ranging between 0 and 1. Our aim is to identify dates d_j, which are relevant for q and minimize any error caused by non-relevant or wrong dates. Our proposal is that the relevance between a (q, d_j) pair is better defined if, instead of just focusing on the self-similarity between the query q and the candidate date d_j, all the information existing between W_j^* and d_j is considered. Thus, we will not only define the similarity between the query words q and the candidate date d_j, but also between each of the most important words $w_{\ell,j} \in W_j^*,\ \ell = 1, \ldots, r_j$ and the respective candidate date d_j. GTE is presented in Eq. (5), where *sim* represents any similarity measure of first or second-order and F an aggregation function of the several $sim(w_{\ell,j}, d_j)$:

$$GTE(q, d_j) = F(sim(w_{\ell,j}, d_j)),\ w_{\ell,j} \in W_j^*. \tag{5}$$

While first-order association measures evaluate the relatedness between two tokens as they co-occur in a given context (e.g. ngram, sentence, paragraph, corpus), second-order co-occurrence measures are based on the principle that two tokens are similar if their corresponding context vectors are also similar. Fig. 3 illustrates two such cases.

In this research, we apply the Infosimba (*IS*) second-order similarity measure, a vector space model supported by corpus-based token correlations proposed by Dias, Alves, and Lopes (2007) as defined in Eq. (6):

$$IS(w_{\ell,j}, d_j) = \frac{\sum_{i\in X}\sum_{j\in Y} S(i,j)}{\left(\begin{array}{c}\sum_{i\in X}\sum_{j\in X} S(i,j)+ \\ \sum_{i\in Y}\sum_{j\in Y} S(i,j)- \\ \sum_{i\in X}\sum_{j\in Y} S(i,j)\end{array}\right)}. \tag{6}$$

IS calculates the correlation between all pairs of two context vectors X and Y, where X is the context vector representation of $w_{\ell,j}$ (i.e., those terms that co-occur with d_j) and Y is the context vector representation of d_j. In particular, $w_{\ell,j}$ and d_j are formed

[1] http://wia.info.unicaen.fr/TokenExtractor/api/Token?query= [April 1st, 2015].

[2] http://dbs.ifi.uni-heidelberg.de/index.php?id=form-downloads [April 1st, 2015].

[3] http://nlp.stanford.edu:8080/sutime/process [April 1st, 2015].

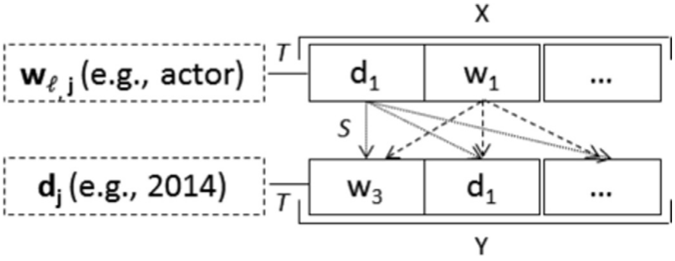


Fig. 4. IS Context vector representation.

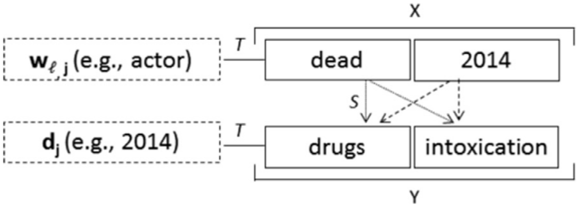


Fig. 5. IS Context vector representation for *actor* and 2014.

by a combination of both terms ($w_1, w_2, .., w_k$) and candidate dates ($d_1, d_2, .., d_t$), having a minimum S similarity value ($S > T$) with $w_{\ell,j}$ and d_j respectively, where S can be any first-order similarity measure such as DICE (Dice, 1945). Similarly, the similarity between each pair of the two context vectors can be determined by any first-order similarity measure $S(i, j)$ relating items (terms or dates) i and j. Fig. 4 illustrates the InfoSimba similarity measure behavior. In the figure, $w_{\ell,j}$ represents one of the several possible words of W_j^*, for example *actor* and d_j one candidate date, for instance *2014* if we consider the query "Philip Seymour Hoffman".

Based on the above representation and on a threshold $T > 0$ we resort to the calculation of the DICE coefficient to determine the eligible context vectors for both *actor* and *2014*. The result will be a vector whose components are arranged in the descending order of the DICE similarity value. For instance, we would obtain for *actor* the vector (*dead, 2014, oscar-winner, most wanted man*) and for *2014* the vector (*drugs, intoxication, overdose, sunday, dead at 46, Manhattan*). After defining the size of the vector we may then determine the final version of the context vectors. For example, for a size set of 2, we would have (*dead, 2014*) as the context vector of *actor* and (*drugs, intoxication*) as the final context vector of *2014*.

IS can now be computed as the corresponding similarity between each pairs of tokens (words *or/and* dates), present in the 2-size context vectors as depicted in Fig. 5.

Specifically, it will compute the level of relatedness between *dead* from the context vector of *actor* and the two other context tokens of *2014* – i.e. *drugs, intoxication* – and then between *2014* and all other context tokens of *2014* and so on and so forth, thus promoting semantic similarity. Note that the similarity between each pair of tokens is again determined by S, which in our example is the DICE coefficient measure. The final score of $IS(actor, 2014)$ which stems from applying Eq. (6) is given by:

$$IS(actor, 2014) = \frac{S(dead, drugs) + S(dead, intoxication) + S(2014, drugs) + S(2014, intoxication)}{\begin{array}{l}(S(dead, dead) + S(dead, 2014) + S(2014, dead) + S(2014, 2014)) + \\ (S(drugs, drugs) + S(drugs, intoxication) + S(intoxication, drugs) + S(intoxication, intoxication)) - \\ (S(dead, drugs) + S(dead, intoxication) + S(2014, drugs) + S(2014, intoxication))\end{array}}$$

Similarly, we should process all the IS similarities between *2014* and the remaining words of $w_{\ell,2014}$. A combination of the several $sim(w_{\ell,j}, d_j)$ is then performed, accordingly to Eq. (6), by an F aggregation function. For the purposes of evaluation, we compared several versions of GTE combined with the IS and the PMI (Church & Hanks, 1990), SCP (Silva, Dias, Guilloré, & Pereira, 1999) and DICE (Dice, 1945) similarity measures, plus different F functions. The best GTE configuration was given by the Median function F and the IS similarity *sim* combined with the DICE similarity measure S. We keep this configuration in this new study. A more thorough discussion of the evaluation methodology can be found in Campos, Dias, Jorge, and Nunes (2012). Each candidate year d_j is then given a temporal similarity value computed by $GTE(q, d_j)$ and stored in a vector called $V_{GTE_{D_S}}$ defined in Eq. (7):

$$V_{GTE_{D_S}} = \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_t \end{bmatrix}, \qquad (7)$$

where T_k, $k = 1, \ldots, t$ represents the weight association between a candidate date d_j and the query q, for the t distinct candidate dates. A web service of GTE is provided[4] so that it can be tested by the research community. The web service returns in XML format, the temporal similarity value calculated between the query and all the candidate dates, together with the corresponding contents where the candidate dates appear.

3.4. GTE-class: date filtering

The final step of our methodology aims to determine whether or not the candidate temporal expressions are relevant to the query. For that, we propose a threshold-based strategy, where a date is considered relevant, if $GTE(q, d_j) \geq \lambda$ and non-relevant otherwise. The final set of m relevant dates for the query q is then defined as D_S^{Rel} :

$$D_S^{Rel} = \left\{d_1^{Rel}, d_2^{Rel}, \ldots, d_m^{Rel}\right\}, \tag{8}$$

where $d_1^{Rel} < d_2^{Rel} < \ldots < d_m^{Rel}$ is an ordered set by relevance. Note that $d_1^{Rel}$and d_m^{Rel} represent the lower and the upper temporal bounds of the query q respectively. Similarly D_{S_i} is decomposed into

$$D_{S_i}^{Rel} = \left\{d_{1,i}^{Rel}, d_{2,i}^{Rel}, \ldots, d_{u,i}^{Rel}\right\}, \tag{9}$$

representing the set of u relevant dates $d_{j,\,i}$ for the query q associated with the web snippet S_i. Based on this, each snippet S_i is no longer represented by a set of candidate temporal expressions as proposed in all reported related works, but by a set of relevant dates. This allows withdrawing non-relevant dates for the query or incorrect ones. As a consequence, we redefine S_i as follows:

$$S_i \rightarrow \left(W_{S_i}, D_{S_i}^{Rel}\right) . \tag{10}$$

Finally, $V_{GTE_{D_s}}$ becomes $V_{GTE_{D_s}^{Rel}}$ such that:

$$V_{GTE_{D_s}^{Rel}} = \begin{bmatrix} GTE_1 \\ GTE_2 \\ \vdots \\ GTE_m \end{bmatrix}, \tag{11}$$

where GTE_k, $k = 1, \ldots, m$ represents the temporal similarity between the date d_j, and the query q, for the m distinct relevant dates and $m \leq t$. In order to determine the best λ, we performed some experiments in Campos et al. (2012), which evidence that best results are obtained with a threshold value of $\lambda = 0.35$ under a stratified 5-fold repeated random sub-sampling validation approach. A description of the GTE-Class demo[5] can be found in our recent work (Campos, Dias, Jorge, & Nunes, 2014a).

4. GTE-rank: temporal re-ranking model

In this section, we describe our temporal re-ranking model, which is the main contribution of this paper. Our aim is to give higher weights to documents having relevant temporal features. Our assumption is that a document should be ranked higher if its contents are topically and temporally related to the query. This is formalized in the principle P1:

P1: The more a given document is correlated to the set of corresponding most relevant words and relevant dates associated with the query, the more the query will be associated with the document.

In order to give users the chance to adjust the temporal and topical parts of the system, we propose a linear model where temporal and topical relevance values are gathered into a single ranking score. GTE-Rank is defined in Eq. (12):

$$GTE - Rank(q, S_i) = \alpha * \sum_{j=1}^{u} GTE\left(q, d_{j,i}^{Re\,l}\right) + (1 - \alpha) * \sum_{h=1}^{k} IS(q, w_{h,i}),\ \alpha\ \in [0, 1], \tag{12}$$

where α is the weight parameter setting the importance of each of the two dimensions, q is the query, $d_{j,i}^{Rel} \in D_{S_i}^{Rel}$, $j = 1, .., u$ is one of the u relevant dates of the snippet S_i and $w_{h,i} \in W_{S_i}$, $h = 1, .., k$ is one of the k most relevant words/multiwords of the snippet S_i.

Central to this ranking function is the computation of two similarities. GTE gives the similarity between the query and each of the relevant dates found in the web snippet. IS gives the similarity between the query and each of the relevant concepts found in the web snippet. Note that one of the advantages of our approach relies precisely on the use of GTE. On the one hand, it enables GTE-Class to filter out the set of all non-relevant or non-date patterns from the input of the ranking module. On the other hand, it allows to dismiss non-relevant dates in the formation of the context concept vectors for the computation of IS, as both the query q and the word $w_{h,\,i}$ are formed by a combination of the best relevant terms and best relevant dates. As a result, we expect

[4] http://wia.info.unicaen.fr/GTEAspNetFlatTempCluster_Server/api/GTE?FilterDates=false&query= [April 1st, 2015]. Note that a query should be appended at the end of the URL. If one wants to get results under a different language other than the default one (en-US), the following code should also be appended together with the desired language. For example, for the Portuguese language we should have "&language=pt-PT".

[5] http://wia.info.unicaen.fr/GTEAspNetFlatTempCluster_Server [April 1st, 2015].

to achieve improved results when compared to state-of-the-art algorithms that simply consider all temporal patterns as equally relevant dates. Below, we formalize the requirement of the ranking function.

R1: S_i is more relevant to q than S_i', if $GTE - Rank(q, S_i) > GTE - Rank(q, S_i')$.

The overall temporal ranking algorithm is formalized below.

Algorithm 1: Temporal ranking.

```
Input: query $q$, weight $\alpha$
S ← GetSnippetsFromSearchEngine($q$)
$D_S$ ← Identify candidate years in S
Compute $GTE(q, d_j),\ j = 1, .., t$, candidate years
$D_S^{Rel}$ ← Determine the $m$ relevant dates by applying $GTE$-Class
Determine $V_{GTE_{D_S}^{Rel}}$
Determine $M_{CT}^{Rel}$
For each $S_i \in S, i = 1,..,n$
    For each $d_j \in D_{S_i}^{Rel},\ j = 1, .., u$
        $GTE + = V_{GTE_{D_S}^{Rel}}\ (q, d_j)$
    For each $w_h \in\ W_{S_i},\ h = 1, .., k$
        $IS + = IS_{M_{CT}^{Rel}}(q, w_h)$
    Compute $GTERank(q,\ S_i) = \alpha * GTE + (1 - \alpha) * IS$
Output: $(q, S_i)$ relevance for each $S_i \in$ S
```

Given a text query q, the algorithm first identifies t candidate years in the set of snippets S. After this, GTE weights the association between the query and the set of t candidate years. The final list of m relevant dates stems from applying GTE-Class. Each of these dates is then stored in the $V_{GTE_{D_S}^{Rel}}$ vector, together with the corresponding association weights. We then determine the M_{CT}^{Rel} matrix which gathers the DICE similarities between "*term*"–"*term*", "*date*"–"*date*" and "*term*"–"*date*". Each snippet S_i is then reordered according to the temporal (GTE) and topical (IS) biased factors. The final temporally biased ranking score is given by the sum of the cumulative values of GTE and IS weighted by $\alpha \in [0, 1]$. In the next section, we define the experimental setup.

5. Experimental setup

In this section, we describe the experimental setup of our approach. In particular, we describe the ground truth dataset, the baseline methods and the evaluation metrics.

5.1. Dataset description

Evaluating time-sensitive information needs is a difficult task since there are no available benchmarks like TREC[6] bringing together queries with an implicit temporal nature and web documents collections whose relevance judgments are made in accordance to the documents contents and not to the timestamp of the document.

Over the years, a few reference collections have been set but they often consist of newswire articles. The TREC 2004 Novelty track[7], for example, created a set of 50 queries of which a few are explicitly tied to a single dated event, despite the query can have multiple temporal instances associated. The system is designed to locate relevant and new information within a set of newswire documents, thus relevance judgments do not take into account any explicit temporal aspect other than novelty.

Another source of TREC queries is based on the TREC 2004 Robust Track[8] news corpus. It gathers some time-sensitive ad hoc queries selected from TREC-{6,7,8} and previous robust tracks, but similar to the novelty dataset it does not determine the correct time of the query nor it produces relevance judgments according to temporal aspects.

The NTCIR-GeoTime[9] challenge, for example, addresses a similar ranking problem to ours but it focuses on queries having temporal and geographic aspects of the form "where and when happened X". Each dataset (NTCIR-8 GeoTime Gey, Larson, Kando, Machado, and Sakai (2010) and NTCIR-9 GeoTime Gey, Larson, Machado, and Yoshioka (2011) consists of 25 queries and a document collection of newswire articles.

More recently another temporal task has been launched. The TREC 2013 and 2014 Temporal Summarization[10] (Guo, Diaz, & Yom-Tov, 2013) task consists of 25 temporal queries and a set of timestamped documents covering the period from October 2011 to April 2013. The goal of the Temporal Summarization track however, is to develop systems which can detect new information related to a developing event over time. This contrasts with our task which is geared towards ranking documents according to the different possible times of the query inferred from the contents of a web document collection.

[6] http://trec.nist.gov [March 1st, 2015].

[7] http://trec.nist.gov/data/t13_novelty.html [April 1st, 2015].

[8] http://trec.nist.gov/data/t13_robust.html [April 1st, 2015].

[9] http://metadata.berkeley.edu/NTCIR-GeoTime/ [April 1st, 2015].

[10] http://www.trec-ts.org [April 1st, 2015].

Table 1
List of queries.

george bush iraq war	avatar movie	tour eiffel	steve jobs	amy winehouse
slumdog millionaire	britney spears	troy davis	waka waka	haiti earthquake
football world cup	justin bieber	adele	nissan juke	marco simoncelli
walt disney company	little fockers	volcano iceland	lena meyer-landrut	ryan dunn
david villa	true grit	bp oil spill	fiat 500	haiti
susan boyle	sherlock holmes	tour de france	lady gaga	katy perry
dacia duster	fernando alonso	david beckham	fukushima	obama
kate nash	osama bin laden	rebecca black		

Finally, the NTCIR-11-Temporalia[11] (Joho, Jatowt, & Blanco, 2014) challenge is a very recent task comprising a document corpus of blog and news sources and a mixed combination of implicit and explicit temporal queries. Temporalia offers two sub-tasks: the temporal classification of queries and a ranking task, where participants are asked to submit the top 100 documents for each query per different temporal class (i.e., past, recency, future and atemporal). Documents relevance is performed according to whether they are relevant to the query or not, with no particular mention to the temporal part of the document, thus constituting a major difference with our approach. In addition, we do not consider explicit temporal queries in our evaluation task nor we distinguish between any different kinds of temporal classes.

Given the inexistence of a TREC-like collection that suits our temporal information retrieval task, we developed a new publicly available dataset[12], gathering 38 implicit temporal queries, 1900 documents and relevance judgments, thus establishing baseline performance for further studies. Note that because of the lack of a public collection that provides temporal relevance for time-sensitive queries it becomes hard to conduct experiments in a larger dataset. Thus, in addition to the experiments carried out we provide a demo search interface (see Section 8), thus providing users and the research community with the possibility of extensively testing the effectiveness of our proposal.

Our collection, named *Query-Snippet Google Insights for Search Bing Ranking dataset* (*QSGisBingRank_DS*) is constructed by running each of the queries into a commercial search engine and by conducting relevance judgments as explained below. In order to gather a representative set of queries, we rely on Google Trends, a Google service which provides users with a visual representation of top and rising searches. We start by considering the 20 queries available on each of the 27 pre-defined categories so as to cover a wide set of domains and reliable conclusions. After removing duplicates and explicit temporal queries, we end up with a set of 450 queries. As we aim to evaluate the topical and temporal relevances of a web snippet, we need to guarantee that the queries selected are topically non-ambiguous[13] and temporal in their purpose, such that no bias is brought into experimental results.

For the first step we have used the Wikipedia disambiguation feature, which helps to understand whether a query has more than one meaning or facet. Final results show that 176 queries are of clear nature, i.e., non-ambiguous. Each clear concept query must then be classified with regard to its temporal nature. For the purpose of judging the set of 176 clear concept queries with regard to their temporality, three human annotators were asked to consider each query, to look at web search results and to classify them as *Temporal* or *ATemporal*. As an alternative to this manual identification, we could have resorted to some temporal categorization strategy, either Wikipedia or snippet-based (Campos et al., 2011). We opted not to use any of these approaches as our intention was to stick as close as possible to the real ground truth, i.e. people, without introducing any potential error into the classification scheme. In the future, we will explore combining our work with a query temporal categorization strategy in order to adopt a more balanced approach between the conceptual and the temporal parts of the ranking model, thus boosting more temporal results when the query is of temporal nature, while promoting more topical ones when the query is deemed to be atemporal.

The final classification comes by majority voting. As such, each query is considered to be ATemporal if it gets at least two votes, while Temporal otherwise. An inter-rater reliability analysis using the Fleiss Kappa statistics (Fleiss, 1971) was then performed to determine consistency among annotators. Results have shown a value of 0.89, thus indicating an almost perfect agreement between the raters. The final set (see Table 1) consists of 38 real-world text clear-concept temporal queries.

Our next step is to obtain a collection of web snippets. For that purpose, we relied on the Bing Search API[14] parameterized with the *en-US* market language parameter to retrieve 50 results per query, which resulted in a set of 1900 web snippets, of which 543 contain year terms (e.g. 2014). Though our algorithm will profit from having access to more snippets, we are limited, as a third party, to retrieve only 50 results per query as set by Bing.

Each (q, S_i) pair was then assigned a relevance label by a human judge on a four-point relevance scale. Our assumption is that users tend to prefer results that carry temporal features, as opposed to those that only have text as shown by Alonso et al. (2011). Using a relevance in topic and time naturally allows to increase the quality of the retrieved results, because documents not fully satisfying both dimensions will tend to appear in lower positions. Under this assumption, a web snippet containing both topical

[11] http://ntcir.nii.ac.jp/Temporalia/NTCIR-11-Temporalia/ [April 1st, 2015].

[12] http://www.ccc.ipt.pt/~ricardo/datasets/QSGisBingRank_DS.html April 1st, 2015].

[13] A query that has a specific meaning and covers a narrow topic usually is a successful search in which the user can find what he/she is looking for in the first page of results, e.g., avatar movie.

[14] https://datamarket.azure.com/dataset/5BA839F1-12CE-4CCE-BF57-A49D98D29A44 [April 1st, 2015].

Table 2
Queries relevance judgments for the two datasets.

Relevance grade	Temp_DS	TempTopic_DS
0	38	417
1	41	213
2	50	662
3	414	608
Total	543	1900

Table 3
Parameter setting. Boldface indicates the best MAP values with regard to the respective parameter.

Parameter	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
TF.IDF - *b*	0.6927	0.6946	0.6947	0.6949	0.6044	0.6949	**0.6953**	0.6951	0.6949	0.6934	0.6919
HLM - λ	0.6928	**0.6928**	0.6926	0.6924	0.6922	0.6920	0.6918	0.6915	0.6910	0.6916	0.6716
PL2 - c	0.6716	0.6700	0.6723	0.6727	0.6731	0.6731	0.6732	0.6730	0.6743	0.6742	**0.6754**
BM25 - *b*	0.6936	0.6961	0.6960	0.6960	0.6953	0.6960	**0.6970**	0.6962	0.6966	0.6940	0.6938

and temporal informations matching the query needs is considered to be extremely relevant and is labelled with a score of 3. It is worth noting that snippets without year temporal information may also get a score of 3 (e.g. "*Amy Winehouse consumed a very large quantity of alcohol before dying at her London home, a pathologist said Wednesday as she declared Winehouse's demise...*" for the query "*Amy Winehouse*") as long as they are topically relevant.

In the opposite direction a web snippet that is not topically, nor temporally relevant, gets a score of 0. Moreover, the simple fact that a result includes a temporal expression is not self-sufficient to get favored. Indeed, web snippets having a year temporal reference may end up getting a score of 0 (e.g. "*©2011 EA Fragrances Co. Britney Spears*$^{\text{TM}}$ *is a trademark licensed to Elizabeth Arden, Inc. by Britney Brands, Inc.*" for the query "*Britney Spears*") whenever they are considered temporally non-relevant.

The labeller was allowed to perform search on the web to get knowledge about the topic and eliminate context factors that might influence a change in his/her judgment. Next, we formed two distinct datasets (see Table 2). The first one, designated *Temp_DS*, comprises only those web snippets having temporal features. *TempTopic_DS*, in turn, includes the set of 50 web snippets retrieved for each query, independently if they contain temporal features or not. Based on these two collections, we can test the GTE-Rank performance in two different scenarios: (1) an exclusively temporal scenario and (2) a scenario involving the combination of topical and temporal relevancies.

5.2. Baseline methods

For the baseline ranking schema, we used the set of results retrieved by the Bing search engine and considered four different baseline ranking models. We believe that comparing our work to a real search engine is an asset of this research. For this, we consider:

1. *Bing*: Bing search engine initial ranking.
2. *Random*: random ranking over Bing search engine results.
3. *AscOrder*: order by ascending date ranking over Bing search engine results.
4. *DescOrder*: order by descending date ranking over Bing search engine results.

As a means to complement our analysis we considered four further purely-relevance baseline ranking approaches. To this end, we employed the Terrier[15] open source search engine to build an index upon the set of web snippets retrieved by Bing and considered a ranked set of documents based on Terrier's implementation. More specifically, we consider:

5. *TF.IDF*: term frequency-inverse document frequency weighting model. TF is given by Robertson's tf (Robertson et al., 1994) and IDF is given by the standard Spärck Jones' idf (Spärck Jones, 1972).
6. *HLM*: Hiemstra language model (Hiemstra, 2001).
7. *PL2*: an advanced divergence from randomness weighting model (Amati, 2003).
8. *BM25*: the BM25 probabilistic model (Spärck Jones, Walker, & Robertson, 2000).

Due to the fact that most of these methods can produce substantially different results lay based on the set of parameters chosen and on the type/length of documents and queries, a significant number of experiments have been conducted to keep things coherent. With this view in mind, *TF.IDF, HLM, PL2* and *BM25* ranking functions have been trained over our collection with a wide number of different settings in order to determine the best learned parameters. The results drawn from the training stage can be observed on Table 3 which lists the MAP (Mean Average Precision) values for both ranking functions under different parameter settings.

[15] http://terrier.org [April 1st, 2015].

From Table 3 we can observe negligible differences between any of the parameter settings with regards to their ranking functions performance. Best values for λ and *c* reveal that HLM and PL2 are respectively in line with the default Terrier's settings. In contrast, experiments with different values of *b* over the training data show that unlike Terrier default value (which is set to *0.75*), $b = 0.6$ is among the best performing figures for both TF.IDF and BM25 ranking methods, though with minimal differences between them. These values will be later on applied all over our experiments.

Finally and in order to compare our approach over related work (Kanhabua & Nørvåg, 2010; Dakka et al., 2012) we consider three additional baselines that make use of temporal signals. More precisely, we consider:

9. *TBM25*: a linear combination between BM25 probabilistic model (Spärck Jones et al., 2000) and a temporal relevance score as in Dakka et al. (2012).
10. *NLM-U*: a linear combination between a topical score and a time score as in Kanhabua and Nørvåg, (2010).
11. *NLM-U_GTE-R*: a linear combination between the NLM-U topical relevance part and the GTE-Rank temporal one (onwards denoted GTE-R for simplicity).

In order to implement each one of these methods we had to adapt our approach to the specificities of both Kanhabua and Nørvåg (2010) and Dakka et al. (2012) solutions, which are metadata-based dependent. To accomplish this, we tailored our approach to a metadata framework using the date appearing in the snippet as the publication date of the document. If there is more than one date, only the most recent will be deemed. Furthermore, just as in our solution, no publication date will be regarded if the web snippets embody no temporal signals. In that case the relevance score will be simply computed with resort to the topical part of the equation. Next, we describe in more detail each one of these methods.

In TBM25 we rely on the work of Dakka et al. (2012) to linearly combine BM25 with a temporal score that depends on modeling time with regard to the relevance of a time point *t* to a query *q*. Following Dakka et al. (2012) we integrate the temporal relevance $p(t|q)$ into the probabilistic BM25 relevance model as in Eq. (13):

$$TBM25(q, S_i) = BM25(q, S_i) + \log \frac{P(t|q)}{1 - P(t|q)} \tag{13}$$

where $BM25(q, S_i)$ is the same as BM25 and $P(t|q)$, with *t* being the publication date of the document S_i, a query likelihood model that is estimated as follows:

$$P(t|q) \propto \prod_{w\ \in q} P(w|t) = \prod_{w\ \in q} \frac{tf\left(w, S^t\right)}{|S^t|} \tag{14}$$

We let *w* denote a word of the query *q* and use $tf(w, S^t)$ to refer to how frequent the term *w* occurs in the set of documents published in time *t*. Likewise, $|S^t|$ denotes the total number of term occurrences in the set of documents published in time *t*.

Another possibility that we explore here is to implement NLM-U approach (Kanhabua & Nørvåg, 2010). Likewise our solution, this work considers a linear combination between a keyword and a temporal score, subject to an α parameter capable of boosting one of the parts in detrimental of the other. NLM-U is defined in Eq. (15):

$$NLMU(q, S_i) = (1 - \alpha) * S'(q, S_i) + \alpha * S''(q^{time}, S_i^t) \tag{15}$$

where $S'(q, S_i)$ is the topical part of the formula and $S''(q^{time}, S_i^t)$ is the temporal one. Further, q^{time} is a set of temporal instances $\{t'_1, \ldots, t'_n\}$ deemed to be relevant to the query and *t* the publication date of the document S_i. Contrary to our work, where only relevant dates to the query are considered, Kanhabua and Nørvåg (2010) assume all the publication dates of the documents to be query relevant.

To determine the topical part of the equation the authors use the Terrier search engine and employ the DFR_BM25 ranking model, proceeding with the normalization of the values by dividing for the maximum keyword score among all the documents. For the temporal part, they resort to the probability of generating the time of the query q^{time} given the associated publication date of the document S_i^t as in Eq. (16):

$$S''(q^{time}, S_i^t) = \frac{1}{|q^{time}|} * \sum_{t'_j \in\ q^{time}} P(t'_j|S_i^t) \tag{16}$$

The probability of generating the time t'_j given the publication date of the document $P(t'_j|S_i^t)$ is then defined by taking uncertainty into account using for that an exponential decay function as in Eq. (17). Intuitively, a document whose publication date is closer to the query time t'_j will be given a higher probability than a document that is far apart from t'_j. The determined times of the query are then assigned a weight $w(t'_j)$ that accounts for their importance using for that the documents reverse ranked number.

$$P(t'_j|S_i^t) = \frac{w\left(t'_j\right)}{\sum_{t'_k \in\ q^{time}} w\left(t'_k\right)} * DecayRate^{\lambda * |t'_j - S_i^t|} \tag{17}$$

Following Kanhabua and Nørvåg (2010) experiments, we use an exponential $DecayRate = 0.5$, $\lambda = 0.5$ and 0.10 for the α parameter.

Our final baseline model is a mixed combination between the topical part of NLM-U and the temporal part of GTE-R. One such combination will enable us to test the behavior of both models under the same circumstances, i.e., on top of the same topical ranking model. As in the original methods, our objective is to bring up documents that are of topical and temporal interests to the query. NLM-U_GTE-R is defined in Eq. (18):

$$NLMU\_GTER(q, S_i) = (1-\alpha) * S'(q, S_i) + \alpha * \sum_{j=1}^{u} GTE\left(q, d_{j,i}^{Rel}\right) \tag{18}$$

where α is the weight parameter setting the importance of each of the two dimensions, $S'(q, S_i)$ is the normalized topical score determined by the DFR_BM25 ranking model, q is the query and $d_{j,i}^{Rel} \in D_{S_i}^{Rel}$, $j = 1, .., u$ is one of the u relevant dates of the snippet S_i.

5.3. Evaluation metrics

To measure how close the generated ranking results are to the ground truth, we used a set of well-known IR metrics commonly used in TREC's evaluations. In particular, we used Precision at k (*P@k*), Recall at k (*R@k*), Mean Average Precision (*MAP*), Mean R-Precision (*MRP*), Mean Reciprocal Rank (*MRR*) and Discounted Cumulative Gain (*DCG@k*). All but the DCG@k are binary metrics, which implies the ground truth to be adapted. Hence, for the grades in Table 2, scores <0, 1> are mapped to the non-relevant label, while scores <2, 3> to the relevant one.

6. Results and discussion

In this section, we describe the set of experiments conducted. Our aim is twofold: (1) to understand the impact of the GTE-Class model (which only considers relevant dates) in terms of our ranking approach; (2) to test the GTE-R ranking effectiveness over the baselines. For this, we consider three experiments.

In our first experiment we aim to test any possible difference in terms of ranking effectiveness that may exist between considering only relevant dates or all the candidate dates. This is one important step of our experimental design as it will enable us to empirically evaluate the merits of the GTE-Class model in terms of our ranking approach. To do so, we test our GTE-R ranking function using two different versions of the GTE temporal similarity measure. One based on $V_{GTE_{D_S}^{Rel}}$, named *GTE-R1*, which only considers relevant dates and another one based on $V_{GTE_{D_S}}$ named *GTE-R2,* which considers all the candidate dates as relevant ones. This first experiment will be conducted on top of the Temp_DS dataset.

In our second experiment, we aim to test the GTE-R ability to pull up relevant documents when compared to baselines making use of temporal signals (i.e., TBM25, NLM-U and NLM-U_GTE-R). To accomplish this, we use the Temp_DS dataset, a strictly temporal collection where all the documents are tagged with a date.

Finally, in our third experiment, we will experimentally evaluate GTE-R effectiveness over a regular web-based collection that combines both temporal and atemporal texts. We rely on the TempTopic_DS dataset and conduct our experiments over all baselines considered. In the upcoming parts, we offer a detailed account of the results obtained for the three experiments.

6.1. Impact analysis of using the GTE-class model on top of the GTE-R method

In this experiment, we compare the GTE-R1 version of our ranking formula (which rests on the GTE-Class date filtering module) against the GTE-R2 version which considers all the candidate dates independently of their relevance to the query (current state-of-the-art). Our aim is to understand how the GTE-Class date filtering module impacts the ranking in terms of the results effectiveness by pulling away from the top temporally detected non-relevant documents. Studying this impact, however, may be compromised if we only restrict our analysis to the top list of the results, as getting a significant number of non-relevant documents on top is unlikely to happen (given that relevant documents are the dominant class). Therefore in order to better understand whether or not there are any differences between GTE-R1 and GTE-R2 and thus to analyze the impact of the GTE-Class in our ranking method, it is also informative to look at the bottom list of the results. We are aware that there is a clear relationship between a system's ability to only rank relevant documents at the top while lowering non-relevant ones to the bottom. Though not directly observable in practice, the tail experiment is, nevertheless, an important step of the evaluation procedure as it allows to understand, in a complementary way, the effectiveness of getting relevant results into the top in contradistinction to non-relevant at the bottom.

With this in mind, we define two different evaluation scenarios:

1. The first one, denoted *Top*, aims to evaluate the ability of the ranking system to gather only relevant documents on the top list of results.
2. The second one, called *Tail*, aims to evaluate the ability of the ranking system to push down all non-relevant documents.

As such, while for the Top approach, MAP measures the average precision over all the queries as regards to the top-k relevant documents, for the Tail one it measures the average precision but this time with regards to the tail-k non-relevant documents. Similarly, MRP considers relevant documents when evaluating the Top scenario by computing the arithmetic mean of all the R-Precision values for the set of all the queries, while non-relevant ones if the Tail approach is being assessed.

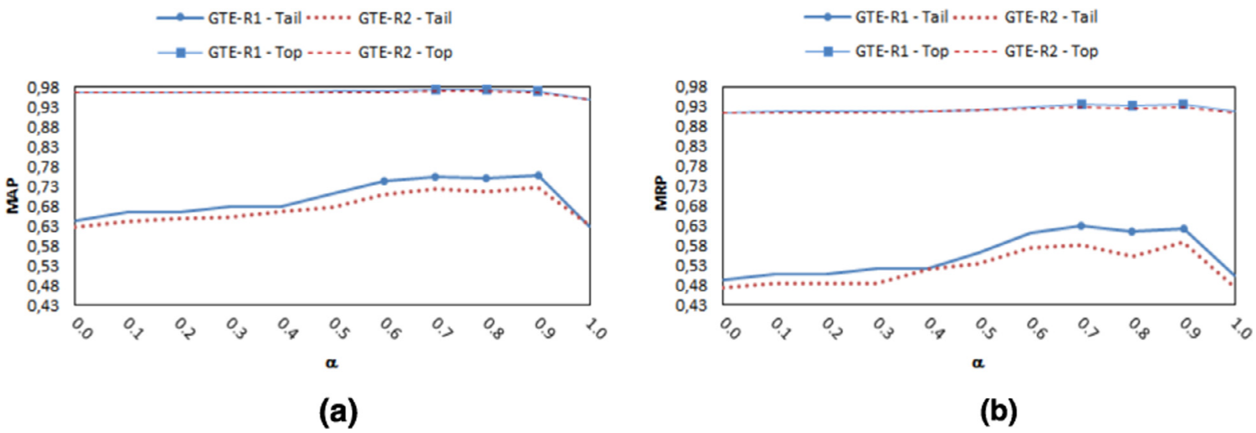


Fig. 6. GTE-R1 vs. GTE-R2. Temp_DS dataset. Top/tail analysis for (a) MAP. (b) MRP.

Table 4
MAP, MRP and $\overline{\text{MRR}}$ results. GTE-R vs. baselines. Temp_DS dataset. Boldface indicates the best value obtained for the respective IR metric.

Method	MAP $\alpha = 0.80$	MRP $\alpha = 0.78$	$\overline{\text{MRR}}$ $\alpha = 0.64$
TBM25	0.886	0.870	0.341
NLM-U	0.885	0.858	0.320
NLM-U_GTE-R	0.952	0.920	0.200
GTE-R	**0.971**	**0.929**	**0.147**

The results obtained over the Temp_DS dataset show that GTE-R1 outperforms GTE-R2 for both scenarios, meaning that our ranking function performs better when the GTE-Class classification module is used. This is clearly illustrated in Fig. 6, where statistically significant improvements (p-Value < 0.05) of the results of GTE-R1 over the GTE-R2 method, using matched paired one-sided t-test, are represented by solid markers. While higher precision scores occur in the Top evaluation scenario, the effect of GTE-R1 is mostly visible in the Tail one. Indeed, if in the case of Top the differences between GTE-R1 and GTE-R2 are minimal (for the reasons previously pointed out), in the case of Tail, GTE-R1 improves MAP and MRP in 0.035 and 0.061, respectively for $\alpha = 0.8$. This was somehow expected as non-relevant dates, to concentrate in the tail-k results, are simply filtered out by GTE-R1, while still considered in the case of GTE-R2. This results confirms that the use of the GTE-Class model enables a significant gain in terms of the results effectiveness. Note however, that the GTE-R2 also performs quite well, as non-relevant dates, though not assigned a value of 0 as in the case of GTE-R1, are given a very low value by the GTE temporal similarity measure, thus contributing to mitigate a greater difference between both methods. A further observation, led us to conclude that the temporal part of our ranking measure has a positive effect in the quality of the retrieved results since they get improved as α increases. This is particularly evident for the Tail approach, with GTE-R1 being improved in 0.122 and 0.129, for MAP and MRP, respectively, when $0.0 \leq \alpha \leq 0.9$. Interestingly, results become worse when changing the value of α to 1.0. We conclude that the best results come from the combination between the temporal factor and the topical one. As for the remaining experiments, we simply rely on GTE-R1 approach (onwards denoted as GTE-R for simplicity) as it has proved to achieve the best performance results.

6.2. Effectiveness evaluation of GTE-R vs. baselines under a temporal collection

In this experiment, we compare the effectiveness of GTE-R against baselines that are also temporally-driven. We build on top of the Temp_DS dataset to experimentally verify the effect of applying GTE-R against TBM25 (Dakka et al., 2012), NLM-U (Kanhabua & Nørvåg, 2010) and NLM-U_GTE-R over a set of documents that are all temporally tagged. In order to accomplish this, we resort to the computation of MAP and MRP, and introduce $\overline{\text{MRR}}$ as the average reciprocal ranks over all the queries at which the first *non-relevant* document is retrieved. This will enable us to understand the effectiveness of our system in pulling away from the top non-relevant documents when compared to state-of-the-art methods. To present the results, we resort to Table 4 which lists the outcomes for all the methods under a 5-fold cross validation setting. We operate by randomly partitioning the set of 38 queries into five folds, the first three containing 8 queries each, and the last two containing 7 queries each. Four folds are used for training, thus selecting the α that maximizes GTE-R and one for testing. This process is then repeated five times, using in each one, a different subset for testing and the remaining one for training. The average performance over the five folds is then used to determine the overall performance of each of the ranking models, GTE-R, TBM25, NLM-U and NLM-U_GTE-R. All the results presented are statistically significant when comparing GTE-R to the corresponding baseline methods with p-Value < 0.05 using the matched paired one-sided t-test. Also recall that as for the case of the $\overline{\text{MRR}}$ metric, the best value is the lowest one.

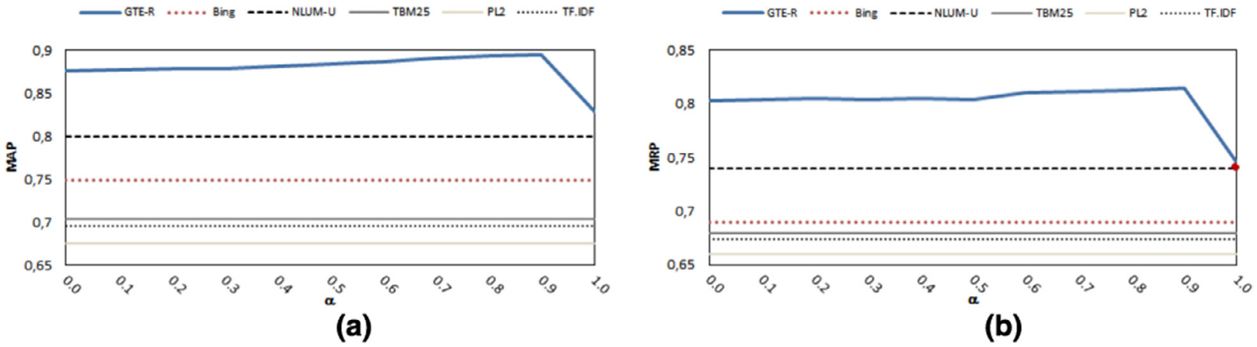


Fig. 7. GTE-R vs. baselines. TempTopic_DS dataset. (a) MAP. (b) MRP.

The results of our empirical evaluation show that there is only a slight difference between applying NLM-U and TBM25, with an advantage for the latter. They further confirm that the NLM-U_GTE-R outperforms NLM-U to a higher extent which highlights the importance of the GTE-R temporal factor. This is particularly evident for $\overline{\text{MRR}}$ with a 0.120 difference, but also for MAP (with a 0.067 difference) and MRP (with a 0.062 difference). From Table 4, we can also show that the topical part of GTE-R algorithm has a positive effect on the effectiveness of the system as GTE-R outperforms NLM-U_GTE-R.

A comparison between our proposal and any of the other two (TMB25 and NLM-U) also shows a notorious difference between both approaches with figures pointing to 0.09 of MAP difference, 0.06 and 0.07 in terms of MRP and 0.19 and 0.17 with regards to $\overline{\text{MRR}}$ for TMB25 and NLM-U respectively. Based on these results, we managed to confirm that the GTE-R proposal performs better than TBM25 and NLM-U. A number of reasons for this can be advanced: (1) we focus on web contents in contrast to the publication date of a document, leveraging all the temporal signals that exist within a text. Thus, if we are faced against two dates we will consider both and not only one; (2) moreover, our system, is able to disregard non-relevant dates detected which again contrasts with related work. This means that we may find a non-relevant date within a document and disregard it, while both TBM25 and NLM-U will consider it to be one important temporal signal of the query. This is clearly demonstrated in the results obtained by the $\overline{\text{MRR}}$ metric; (3) finally, we assign different values for any different date found in the text by taking into account their relevance with the query, thus providing a more comprehensive definition of temporal uncertainty. This means that a document containing a date deemed to be highly relevant to the query will be given a larger weight, while a document that includes a not so relevant or even non-relevant date will be assigned a lower or no score at all. A major difference when compared to the related work. Indeed, TBM25 does not consider uncertainty, while NLM-U does not take into account the fact that a document may eventually refer to a different time point than that of the publication date of the document, as only a comparison between the date of the query and the latter will be made.

Note that none of these results reflect the fact that the related work is considering as the publication date of the document, the temporal signals directly obtained from the text itself. Thus, differences between both approaches will eventually end up being higher in case of a real search scenario where both proposals will be using the publication date of the document, instead of a temporal reference extracted from the text.

6.3. Effectiveness evaluation of GTE-R vs. baselines under a temporal and topic collection

We now test the performance of GTE-R on a collection that also includes atemporal web snippets, i.e. texts which do not include any temporal features. In order to do this, we resort to the unreduced TempTopic_DS dataset and conduct our experiments on top 1900 web snippets collected. We start by considering the difference between the GTE-R results when varying α from 0.0 to 1.0. An overall analysis (see Fig. 7) shows that GTE-R improves as α increases and outperforms the selected baselines when α varies from 0.0 to 1.0, which is consistent with the results obtained so far. Note that for ease of comparison only a few baselines have been included together with GTE-R. We refer to Bing, PL2, TF.IDF, and to NLM-U and TBM25 as temporal approaches. Statistical significance (p-Value < 0.05) of the results is represented by the absence of a solid marker in each of the corresponding lines, when comparing GTE-R over each baseline method.

A complement of this analysis is given in Fig. 8 for the $\overline{\text{MRR}}$ metric. The snapshot indicates that GTE-R achieves again the best score when compared to baselines on pushing down non-relevant documents when α varies from 0.0 to 1.0. This attests to the ability of our system in warding off non-relevant snippets from the top of the list. One reason for this might be due to the use of the GTE-Class which makes it possible for q and $w_{h,\,i}$ to be defined as two context vectors consisting of a combination between relevant words and relevant dates, instead of non-relevant ones.

A summary of the best results for the different baseline measures is given in Table 5 where α has been learned by operating a 5-fold cross validation scheme as in the previous experiment. From this table, we can note that the best values occur for GTE-R proving that GTE-R is capable of obtaining a good performance even over atemporal texts when α is trained. This was expected and confirmed the results obtained in the previous experiment for Temp_DS thereby providing support to the claim that our

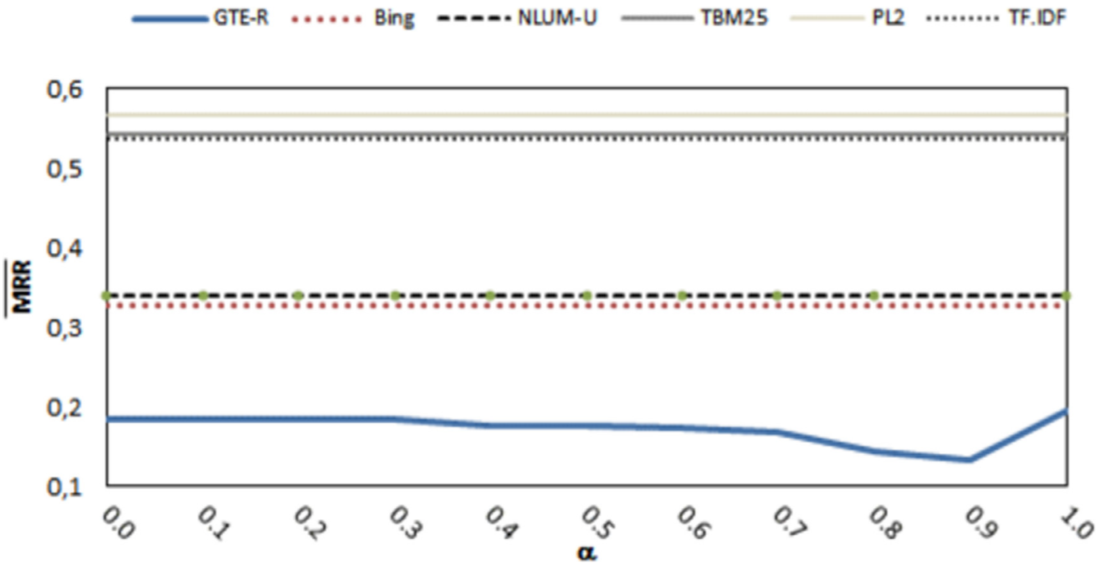


Fig. 8. $\overline{MRR}$. GTE-R vs. baselines. TempTopic_DS dataset.

Table 5
P@k, NDCG@k, MAP, MRP and $\overline{MRR}$ results. GTE-R vs. baselines. TempTopic_DS dataset. All the results are statistically significant when comparing GTE-R against the baseline methods with p-Value $<$ 0.05 using the matched paired one-sided t-test.

Method	P@5 $\alpha = 0.86$	P@10 $\alpha = 0.86$	P@20 $\alpha = 0.78$	NDCG@5 $\alpha = 0.90$	NDCG@10 $\alpha = 0.90$	NDCG@20 $\alpha = 0.90$	MAP $\alpha = 0.88$	MRP $\alpha = 0.88$	$\overline{MRR}$ $\alpha = 0.88$
GTE-R	0.959	0.945	0.890	0.979	0.975	0.968	0.900	0.820	0.118
Bing	0.764	0.769	0.706	0.933	0.895	0.878	0.734	0.682	0.342
Random	0.677	0.648	0.655	0.831	0.793	0.774	0.681	0.648	0.497
AscOrder	0.870	0.843	0.784	0.921	0.927	0.927	0.790	0.729	0.338
DescOrder	0.798	0.821	0.781	0.889	0.891	0.897	0.777	0.729	0.440
TF.IDF	0.665	0.644	0.661	0.799	0.774	0.761	0.687	0.666	0.521
HLM	0.650	0.659	0.666	0.790	0.746	0.752	0.684	0.663	0.585
PL2	0.643	0.632	0.645	0.788	0.750	0.748	0.667	0.643	0.609
BM25	0.586	0.624	0.611	0.763	0.725	0.728	0.665	0.674	0.670
TBM25	0.576	0.620	0.618	0.795	0.726	0.728	0.667	0.675	0.670
NLM-U	0.833	0.840	0.785	0.922	0.905	0.911	0.788	0.729	0.394
NLM-U_GTE-R	0.928	0.882	0.788	0.966	0.965	0.956	0.818	0.736	0.193

approach outperforms related work. An in-depth analysis of the “whys” behind these results has already been presented and discussed in our previous experiment.

By looking at the table, we can also observe that NLM-U_GTE-R is able to achieve the second best result, which is again in line with the results obtained under the Temp_DS dataset. Unsurprisingly, we also found the results of AscOrder to be highly effective. One reason for this is that this method pulls up to the top all the web snippets having dates, which will naturally result in an enhanced performance. Regardless of this, GTE-R still significantly outperforms AscOrder by 0.11 in MAP, 0.09 in MRP, 0.22 in $\overline{MRR}$, 0.05 in NDCG@10, 0.10 in P@10. We conclude that simply using a system that pushes to the top documents incorporating possible temporal features may not be sufficient to achieve a good performance as it is subject to a high degree of randomness. On the one hand, some of the documents will still be relevant to the query although not incorporating any temporal feature. On the other hand, there will be some documents which, although including a temporal pattern, may not be as relevant as those that do not include any date at all (e.g. “*Office 2007*”).

Furthermore, we should call attention to the fact that there is only a slight difference between applying BM25 and TBM25 meaning that the introduction of the TBM25 temporal factor on top of the BM25 ranking algorithm is not enough to produce meaningful results. Two reasons for this can be advanced. First, this may be due to temporal uncertainty absence reasons as no temporal weight will be given in case the query word co-occurs with other temporal instances found in the text than those of the document publication time. Second, the fact that no query dating process is considered. This means that instead of using the set of possible relevant dates to the query to boost the temporal part of the results, only information extracted from the document publication time will be taken into account. A significant limitation that contrasts with both GTE-R and NLM-U proposals.

To further complement this analysis and to compute the normalized distances between each ranking method and the TempTopic_DS ground-truth dataset, we used two widely known metrics, Kendall Tau (Kendall, 1938) and Spearman footrule (Spearman, 1987). While Kendall tau ranking distance is a metric that counts the number of pairwise disagreements between

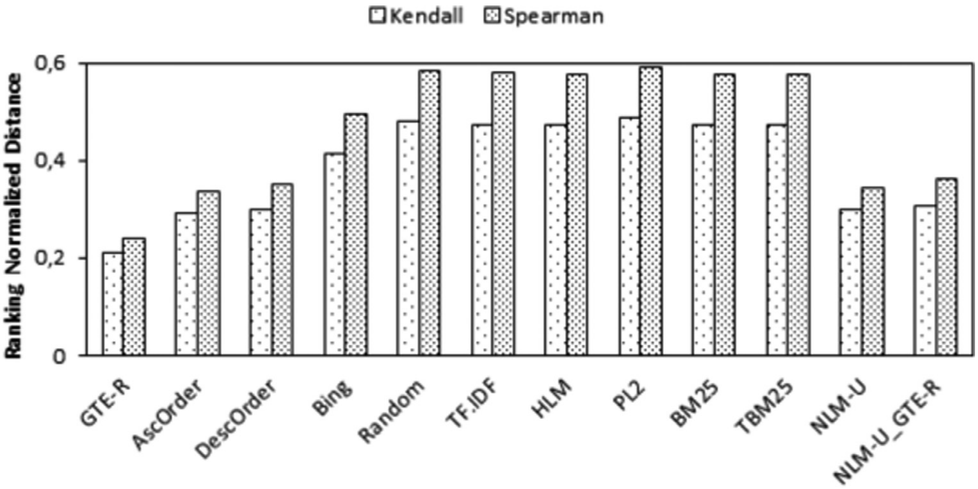


Fig. 9. Kendall Tau and Spearman footrule distance between the GTE-R and the different ranking methods. TempTopic_DS ground-truth dataset ($\alpha = 0.9$).

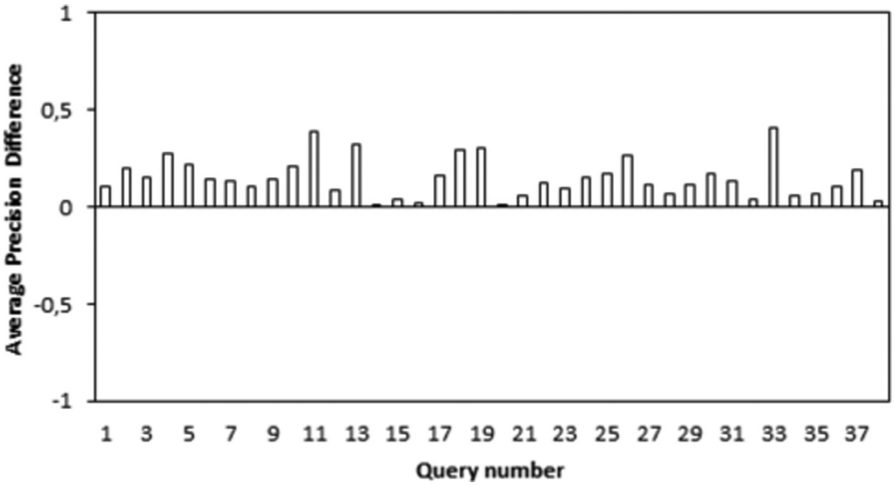


Fig. 10. AP difference histogram for the 38 queries. GTE-R ($\alpha = 0.9$) vs. baselines. TempTopic_DS dataset.

two ranking lists, i.e. the number of pairs of items that are ordered inversely with regard to one another in the two lists, Spearman footrule ranking distance computes the sum of the absolute difference for each item in the two lists. Both distance measures have been extensively used to compare the results returned by two different ranking lists (Kumar & Vassilvitskii, 2010) and thus are a good way to understand how GTE-R behaves as to remaining algorithms.

Fig. 9 shows the results obtained for the two metrics (values were normalized and are within [0,1]). Larger values indicate a higher disagreement between any two ranking lists. As it can be observed, the GTE-R algorithm with a Kendall Tau distance of only 0.210 to the ground truth and a Spearman Footrule distance of 0.238 largely outperforms any of the baseline measures. This further strengthens the results of our empirical evaluation which suggests that GTE-R is able to outperform related work methods.

In what follows, we explore the results of P@k on a per-query basis. For that, we use average precision difference histograms for each query (see Fig. 10), computing the difference between the average precision of GTE-R and the median of the average precisions of the 12 ranking models. The results obtained show that the proposed ranking mechanism outperforms baseline methods as all the queries achieved a positive precision.

Finally, Fig. 11 shows Precision/Recall curves for GTE-R and a considerable number of baselines. For this, we followed the interpolation method suggested by Croft, Metzler, and Strohman (2009) to compute precision values for all standard recall levels (from 0.0 to 1.0). The precision P at any standard recall level R is defined in Eq. (19), where S is the set of observed (R, P) points for a given query, i.e., the set of Recall/Precision values for each retrieved document.

$$P(R) = \max\{P' : R' \geq R \land (R', P') \in S\}. \tag{19}$$

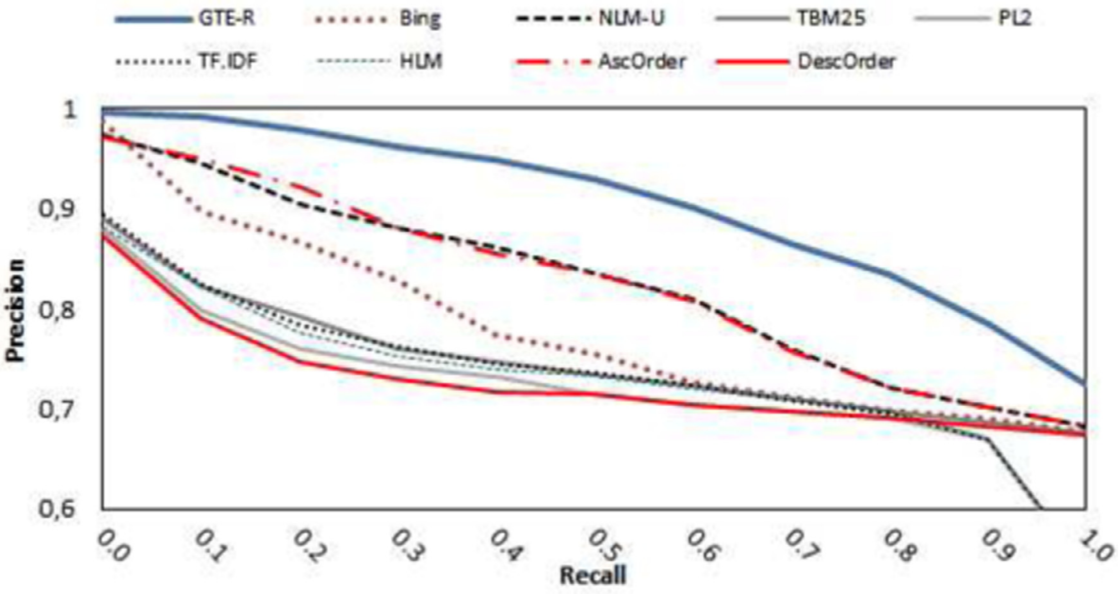


Fig. 11. Average recall–precision curve. GTE-R ($\alpha = 0.9$) vs. baselines. TempTopic_DS.

While GTE-R performs well for all the recall levels, its performance naturally decreases as it approaches 1.0 of recall. This is particularly observable when recall shifts from 0.5 to 1.0, which suggests that while some of the non-relevant documents are still mistakenly dispersed in higher up positions, some of the relevant ones are still incorrectly placed in the lower part of the results. Two reasons for this can be advanced. Firstly, there are some documents for which a date is not relevant, yet GTE-Class defines it as such, or the opposite, i.e. documents for which a date is relevant, yet GTE-Class defines it as non-relevant. With regard to this, it is important to note that the GTE-Class aims to date implicit temporal queries and not to evaluate the relevance of dates within documents. Thus, it can determine that the date "*2011*" is a relevant year for the query "*Steve Jobs*", but it cannot evaluate whether this date is relevant within a snippet (e.g. "*Steve Jobs – February 24, 1955–October 5, 2011*") and non-relevant within another one (e.g. "*Steve Jobs fielded some customer service requests updated: Wed Nov 23 2011 05:51:00*"). This issue must clearly be improved in future research. Secondly, there are some texts, which tend to be pulled up, even if they are not temporally related with the query. This happens with texts embodying words that though relevant with the query refer to a facet of it. One possible way to overcome this is to apply a temporal clustering approach that is able, not only to detect the temporal issues of the query, but also faceted query topics. This is again another important issue for future work and can be handled with multifaceted state-of-the-art clustering algorithms such as proposed in Scaiella, Ferragina, Marino, and Ciaramita (2012). Notwithstanding the limitations laid out above, GTE-R still outperforms the second best approach (i.e., NLM-U) in 0.042 when the recall level equals to 1.0.

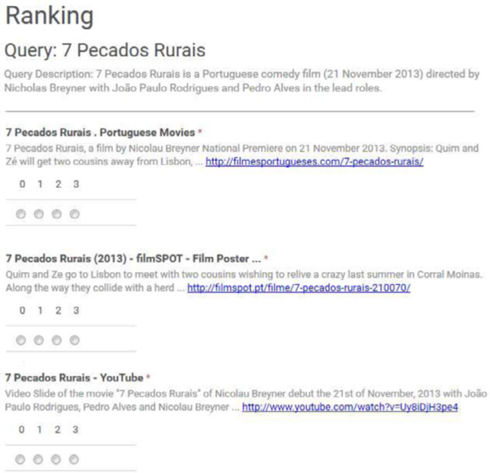
7. User study

In order to measure the effectiveness of our retrieval system on a real-web user environment, we run a crowdsourcing experiment. Our objective is to compare the results of our approach against state-of-the-art methods and to prove the language-independence feature of our system with respect to user judgment. With this in mind, we devised a new dataset, named *Query-Snippet Portuguese Google Trends Bing Ranking dataset* (*QSPTGtBingRank_DS*[16]), consisting of 25 time sensitive queries selected from the archives of the 2012–2014 Portuguese Google Trends[17], a public facility of Google search engine which has been widely used for research tasks. A list of all the queries, from the fields of politics, business, national and international figures, health, sports, gadgets and movies is provided in Table 6. Other queries were simply not eligible due to a number of different factors. For example, some of the queries or its corresponding results are in English as is the case of "iPhone6" or "Cristiano Ronaldo" which albeit in Portuguese retrieves a large number of results in English and Spanish. A few others are already tagged with a temporal feature ("World Cup 2014") or are simply temporally ignorant (e.g., "how to make money"). There are also quite a lot of queries where the information need is topically ambiguous (e.g., "dancing days" which may either refer to a soap or a song) plus a number of different query categories which vary consecutively year-on-year making it difficult to hold a consistent selection (for example we may have the gadgets category appearing on 2014 but not in 2012 or 2013).

[16] http://www.ccc.ipt.pt/~ricardo/datasets/QSPTGtBingRank_DS.html [April 1st, 2015].
[17] http://www.google.com/trends/topcharts?hl=pt#date=2014&geo=PT [April 1st, 2015].

Table 6
List of queries.

surto de legionella	rodrigo menezes	maya gabeira	sara sampaio	telexfree
antónio josé seguro	judite sousa	michele brito	érica fontes	a idade do gelo 4
francisco louçã	manuel forjaz	garret mcnamara	frozen	bq aquaris 5
pedro passos coelho	bernardo sasseti	duquesa de alba	7 pecados rurais	eusébio da silva
paulo portas	margarida marante	josé wilker	a gaiola dourada	o céu existe mesmo



Ranking

Query: 7 Pecados Rurais

Query Description: 7 Pecados Rurais is a Portuguese comedy film (21 November 2013) directed by Nicholas Breyner with João Paulo Rodrigues and Pedro Alves in the lead roles.

7 Pecados Rurais . Portuguese Movies *

7 Pecados Rurais, a film by Nicolau Breyner National Premiere on 21 November 2013. Synopsis: Quim and Zé will get two cousins away from Lisbon, ... http://filmesportugueses.com/7-pecados-rurais/

0 1 2 3
○ ○ ○ ○

7 Pecados Rurais (2013) - filmSPOT - Film Poster ... *

Quim and Ze go to Lisbon to meet with two cousins wishing to relive a crazy last summer in Corral Moinas. Along the way they collide with a herd ... http://filmspot.pt/filme/7-pecados-rurais-210070/

0 1 2 3
○ ○ ○ ○

7 Pecados Rurais - YouTube *

Video Slide of the movie "7 Pecados Rurais" of Nicolau Breyner debut the 21st of November, 2013 with João Paulo Rodrigues, Pedro Alves and Nicolau Breyner ... http://www.youtube.com/watch?v=Uy8iDjH3pe4

0 1 2 3
○ ○ ○ ○

Fig. 12. Google forms human intelligent task. Top-3 results. Here translated to English.

Each query was issued on Bing search engine through the Bing Search API parameterized with the *pt-PT* market language parameter to retrieve 50 results per query, which resulted in a set of 1250 web snippets. To build our test collection, we follow a pooling strategy (Spärck Jones & Bates, 1977; Spärck Jones & Van Rijsbergen, 1975) where only a fraction of the documents consisting of the highest ranking results is considered for assessment thus avoiding the labor of judging the entire dataset. The set of relevance assessments is then used to evaluate all systems.

The selection of the systems and the number of results to retrieve per system stems as a tradeoff between the sample size and the time requirements. With this in mind, we consider three systems to compare, i.e., GTE-R, NLM-U and Bing, where the GTE-R ($\alpha = 0.9$) is our proposal with the best results on the experimental section, NLM-U the second best temporal approach and Bing the best non-temporal one. Note that AscOrder has not been considered as it presents similar results to NLM-U. In addition, we rely on the top-50 ranked documents gathered for each query topic and retrieval system to select the top-10 documents into a pool for assessment, thus guaranteeing that each system contributes with the same exact number of documents. We build upon the work of Zobek (1998) who concluded that using a pool depth of 10 can produce reliable results.

Our pooling strategy resulted in a total of 475 distinct (q, S_i) pairs. Each of these (q, S_i) pairs was then assessed by 33 workers yielding 15,675 (q, S_i) total assessments in a lengthy, labor-intensive task. A mixed combination of research students from our lab (16) and a social list of contacts (17) comprise our list of workers. Most of the workers were national but there were also some multinational workers with high advanced language proficiency. Relevance assessments were collected using Google forms. Workers were provided with a short indicative description of the query to ensure they are familiar with the search topic. Each worker was asked to evaluate the relevance of each (q, S_i) pair under a four-point relevance scale in the same manner as in Section 5.1, where "0" corresponds to a non-relevant result (it does not contain any relevant information), "1" means a marginally relevant one (where the information provided tends to be of low quality, either inaccurate or only partially relevant in a way that it will hardly contribute to enhance the workers knowledge), "2" a relevant document (where the worker is expected to gain some new insights though there might be some better documents) and "3" a highly relevant source (which presents exhaustive/complete information about the topic in a way that is likely to be clicked). Relevance criteria was carefully explained to the workers so that a distinction between documents rich in topical and temporal information (highly relevant and relevant documents) and poor in both strands (non-relevant and marginally relevant documents) is set forth as strictly as possible. The assessments were performed on March 2015 and did not involve any payment. Each worker spent one hour and a half on average to complete the task.

To make sure the goal of the task was fully understand workers were given guidelines of their work and an introductory description of the objectives of the experiment. Fig. 12 shows a screenshot (here translated to English) of the human intelligent task in Google forms for the query "7 pecados rurais" a Portuguese comedy film. Workers are asked to consider the query, to look at the description and to the web search results, and to classify them according to a four-point relevance scale basis. We assume that workers already have an idea about the topic (as given by the description) and that they want to answer a general topical and temporal information need. For example, a web snippet that tells something about the script of the movie, but does not

Table 7
Fleiss' Kappa statistics for each of the three ranking systems.

Method	Kappa	Percentage of overall agreement
GTE-R	0.466	0.733
Bing	0.296	0.648
NLM-U	0.392	0.696

Table 8
Distribution of the documents based on the number of workers voting on non-relevance.

Method	0	[1,6]	[7,13]	[14,20]	[21,27]	[28,33]	# Docs
GTE-R	7	155	69	17	2	0	250
Bing	3	84	68	61	27	7	250
NLM-U	6	119	66	41	14	4	250

refer to its release date, is topically relevant but does not offer the user valuable temporal knowledge that would help him/her to scope the topic into the temporal context space. Naturally, Facebook, LinkedIn, Twitter, YouTube and other official-like web pages would also be considered relevant informative sources despite not being tagged with temporal references. Workers must also have an open-mind not to get stuck on the topic description as several other co-related informations (be it topical or temporal) might also be relevant. One illustrative example of this is the query "7 pecados rurais" which though inherently associated with the "2013" released date, might also find other relevant results from "2015" if a sequel of the movie is forecasted to that date. This should be evaluated on a point-to-point basis.

In order to validate the results of our crowdsourcing experiment, we conducted a set of statistical measurements. We start by studying the consistence of the workers in expressing the same judgment when evaluating the same (q, S_i) pair. To this end, we compute Fleiss' Kappa statistics (Fleiss, 1971). Despite being a fairly straightforward task, we do not expect to reach a high consensus between the workers due to a certain degree of subjectivity which involves this task. The obtained results confirm our assumption by pointing to a 0.324 kappa value and 0.662 of overall agreement, which can be seen as a fair agreement between workers. To better understand these results, we proceed by calculating Fleiss Kappa statistics for each of the three ranking systems. That is, instead of considering the set of distinct 475 (q, S_i) pairs, we treat each system individually by restricting to the set of corresponding 250 (q, S_i) pairs, i.e., 10 results per each of the 25 queries. Our experimental results enable us to conclude that a large majority of the workers tend to agree between them when it comes to labeling the results of our ranking proposal GTE-R. This contrasts with the results of Bing and NLM-U ranking systems where disagreements tend to occur more frequently. A detailed list of the Kappa values can be seen in Table 7.

In order to understand the impact of these disagreements, we aggregate the number of workers disagreeing on the relevance of a document into the following set of intervals: [1,6]; [7,13]; [14,20]; [21,27]; [28,33]. For example the interval [1,6] counts the number of documents (among the 250 returned) deemed to be non-relevant for [1,6] workers out of the total number of 33 workers. Table 8 shows a summary of the results where "0" stands for those cases where no disagreements between the works occur, i.e., all the workers agree on the relevance of a document. As previously indicated, the impact of the disagreements is mostly felt within the Bing and NLM-U proposals. While most of the disagreements tend to be expressed by a minority of workers ranging from 1 to 6, a considerable number of documents retrieved are yet considered to be non-relevant by a large majority of the annotators, particularly for the state-of-the-art methods. Indeed, 61 documents of Bing were deemed to be non-relevant by a large number of workers comprised between [14,20], 41 belonging to NLM-U, but only 17 pertaining for GTE-R, a cross-cutting issue among all the intervals. This difference turns out to be even more evident if we consider an aggregated interval comprised between [14,33]. In this case, we get 95 documents of Bing deemed to be non-relevant by a large number of the workers, 59 for NLM-U and only a small number of 19 documents for GTE-R, which highlights the good behavior of our method. Another important evidence comes from the fact that the NLM-U method seems to present better results than the Bing algorithm. However, this will not be confirmed in further experiments.

In order to deeply understand the workers' satisfaction with regards to the retrieved results, we also study the distribution of the relevance grades for each of the three ranking systems (see Fig. 13) with respect to the whole set of 33 × 475 grades given. Once again Bing gets more non-relevant labels than any other system in line with the previous experiment. These results further confirm that our system gathers a large number of higher relevance scores when compared to remaining approaches meaning that, in general, workers tend to prefer our results rather than those of related work approaches. This will be confirmed in the following experiment.

In our final evaluation, we measure the effectiveness of the different approaches under comparison by resorting to Precision at k (*P@k*) and Normalized Discounted Cumulative Gain (*NDCG@k*) measures. Relevance grades were once again adapted in the case of *P@k* to a binary labeling scheme in order to answer its binary structure. The threshold to define a document as relevant lies between potentially useful documents (score 2) and relevant but useless (score 1). On these grounds, labels of the form $<0,1>$ are defined as non-relevant and thus re-scaled to 0, while labels of the form $<2,3>$ are interpreted as a relevant grade and thus given a score of 1. *NDCG@k* in turn keeps its 4-point relevance scale structure, which allows relevance scores to be weighted

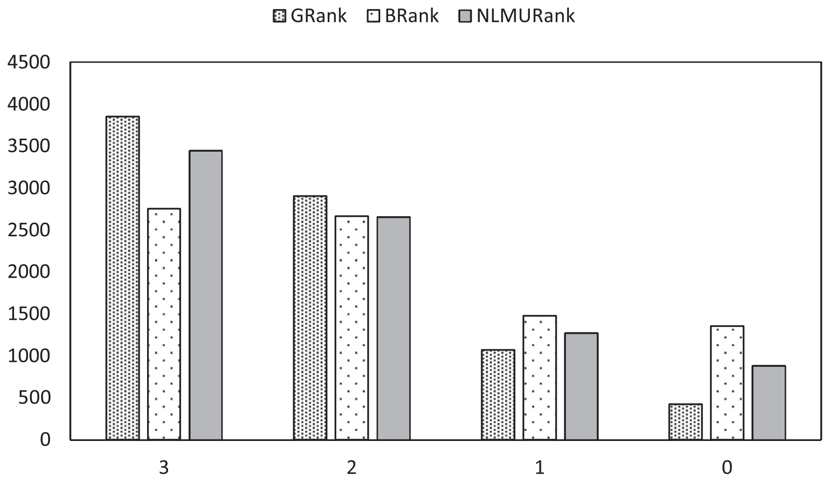


Fig. 13. Relevance grades distributed by ranking system.

Table 9
P@k and NDCG@k results.

Method	P@5	P@10	NDCG@5	NDCG@10
GTE-R	0.877	0.819	0.851	0.933
Bing	0.706	0.657	0.766	0.887
NLM-U	0.788	0.739	0.724	0.858

differently. The various judgements for each pair are then aggregated as an average of the relevance scores determined for each worker. A summary of the results is presented in Table 9.

The results further confirm that the GTE-R algorithm consistently outperforms each of the two other ranking systems over the four different measures. This is particularly evident for P@10 where a considerable difference of 0.171 between GTE-R and Bing can be observed. Best results however occur for NDCG@10 with a 0.933 value. Our results also show a notorious difference between NDCG@10 (0.933) and P@10 (0.819) values, for GTE-R, which further confirm that non-relevant results tend to occur on lower rank positions. As previously pointed out the Bing algorithm outperforms the NLM-U for both NDCG@5 and NDCG@10. However, NLM-U performs better than its counterpart both for P@5 as well as for P@10 which means that, though Bing may comprise more non-relevant documents yet it has the ability of pushing them further down to the end of the list.

The provided evidence is in line with the results previously obtained in Section 6.3 (see Table 5) and supports the claim that our approach is significantly better than the related work over different languages. This is not surprising since our approach rests on the calculation of frequencies of tokens/words/multiwords, without any kind of knowledge-based language dependence.

8. Search interface

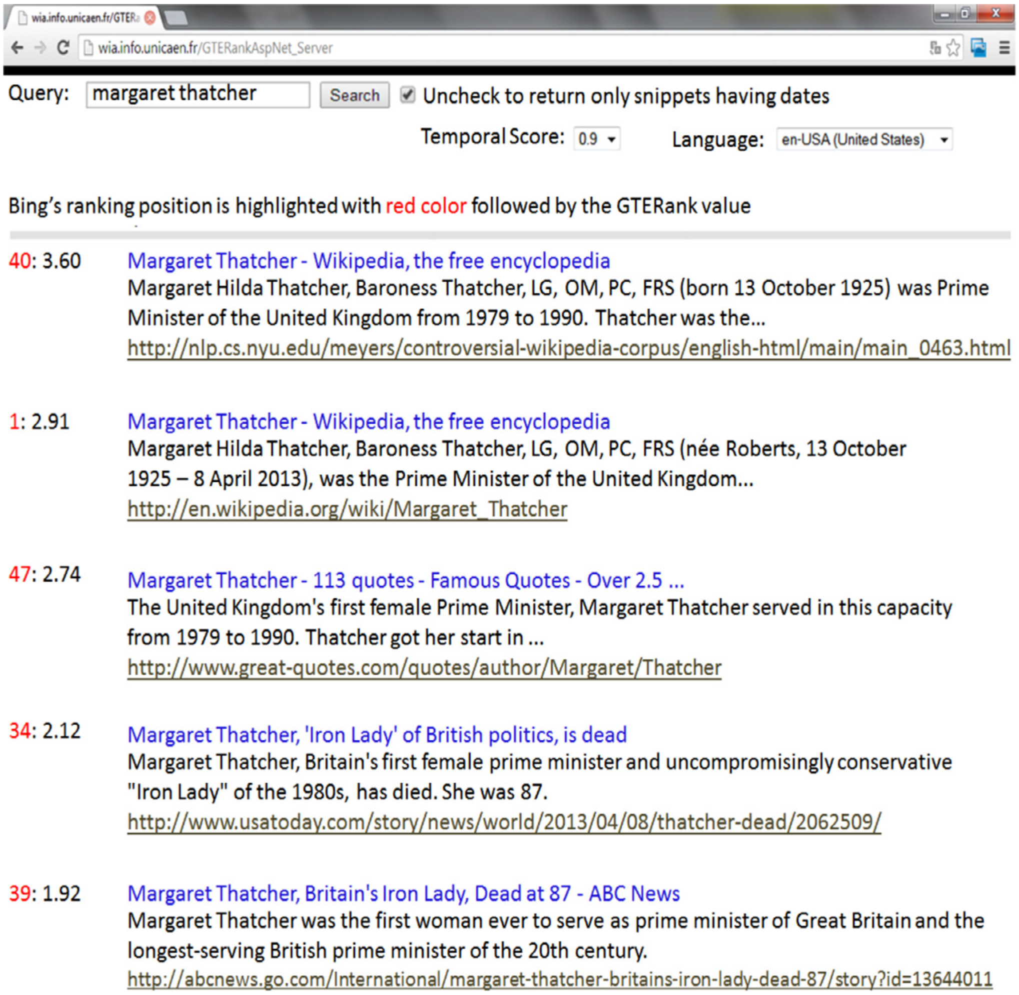
As a result of our research, we publicly provide a set of web services and an online demo (http://wia.info.unicaen.fr/GTERankAspNet_Server) to be tested by the research community. By doing this, we offer users the chance to try their own query examples. The GTE-R web application (Campos, Dias, Jorge, & Nunes, 2014c) can easily be tested online (limited to 5000 queries per month as set by Bing search engine). Topical and temporal expressions detected in documents are appropriately encoded for efficient look-ups to determine relevant documents to queries with a topical and temporal information need as quickly as possible. Although the main motivation of our work is focused on queries with temporal nature, the implemented prototype allows the execution of any query including non-temporal ones. Since our system does not pose any constraint in terms of language (as far Occidental languages is concerned), domain or time period covered, users can issue queries without any kind of restriction, ranging from the business domain (e.g. "*iPad*"), to cinema (e.g. "*true grit*"), politics (e.g. "*Margaret Thatcher*"), natural disasters (e.g. "*Haiti earthquake*") or musical topics (e.g. "*Radiohead*"), to cite just a few. In the following, we provide a detailed account of the web services made available. Note that, in order to work, each web service should be added a query at the end of the URL.

- GTE-R_1 [18] returns, in XML format, the set of fifty re-ranked web snippets.
- GTE-R_2 [19] returns, in XML format, a filter of the re-ranked web snippets, i.e., those containing only relevant dates[20].

[18] http://wia.info.unicaen.fr/GTERankAspNet_Server/api/GTERank?AllSnippets=true&query= [April 1st, 2015].

[19] http://wia.info.unicaen.fr/GTERankAspNet_Server/api/GTERank?AllSnippets=false&query= [April 1st, 2015].

[20] Note that a query should be appended at the end of the URL. If one wants to get results under a different language other than the default one (en-US), the following code should also be appended together with the desired language. For example, for the Portuguese language we should have "&language=pt-PT".



Query: margaret thatcher ☑ Uncheck to return only snippets having dates

Temporal Score: 0.9 Language: en-USA (United States)

Bing's ranking position is highlighted with red color followed by the GTERank value

40: 3.60 Margaret Thatcher - Wikipedia, the free encyclopedia
Margaret Hilda Thatcher, Baroness Thatcher, LG, OM, PC, FRS (born 13 October 1925) was Prime Minister of the United Kingdom from 1979 to 1990. Thatcher was the...
http://nlp.cs.nyu.edu/meyers/controversial-wikipedia-corpus/english-html/main/main_0463.html

1: 2.91 Margaret Thatcher - Wikipedia, the free encyclopedia
Margaret Hilda Thatcher, Baroness Thatcher, LG, OM, PC, FRS (née Roberts, 13 October 1925 – 8 April 2013), was the Prime Minister of the United Kingdom...
http://en.wikipedia.org/wiki/Margaret_Thatcher

47: 2.74 Margaret Thatcher - 113 quotes - Famous Quotes - Over 2.5 ...
The United Kingdom's first female Prime Minister, Margaret Thatcher served in this capacity from 1979 to 1990. Thatcher got her start in ...
http://www.great-quotes.com/quotes/author/Margaret/Thatcher

34: 2.12 Margaret Thatcher, 'Iron Lady' of British politics, is dead
Margaret Thatcher, Britain's first female prime minister and uncompromisingly conservative "Iron Lady" of the 1980s, has died. She was 87.
http://www.usatoday.com/story/news/world/2013/04/08/thatcher-dead/2062509/

39: 1.92 Margaret Thatcher, Britain's Iron Lady, Dead at 87 - ABC News
Margaret Thatcher was the first woman ever to serve as prime minister of Great Britain and the longest-serving British prime minister of the 20th century.
http://abcnews.go.com/International/margaret-thatcher-britains-iron-lady-dead-87/story?id=13644011

Fig. 14. GTE-R interface for the query "margaret thatcher". Top 5 results. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

The graphical search interface is public and helps users searching for information of a given topic through time without any temporal or language constraint. The implemented version is designed to demonstrate the GTE-R functionality, thus concerns of design nature where not taken into account. In response to a query submitted in a search box, GTE-R displays a set of ranked web snippets on the fly.

We offer two types of retrieval: one that returns only web snippets having dates and one that returns the set of all the 50 web snippets, whether or not they have dates. In addition, we give users the chance to adjust the temporal and topical parts of the system. Through an interactive browsing tuning parameter, the user is thus able to define the importance of the two dimensions. α is currently preset to 0.9 as GTE-R has achieved the best performance with this value in the experiments carried out. Each web snippet is also assigned a relevance ranking value reflecting its topical and temporal similarity with the user's query. This value is positioned in front of the number in red color, which defines the ranking position initially obtained by Bing search engine. The user can also choose from a list of languages from which he/she would like to run the query. An illustration of this can be seen in Fig. 14 for the query "Margaret Thatcher".

Following, we present a few examples which illustrate interesting facts that cannot be directly inferred by means of quantitative evaluation metrics. The queries selected ("Margaret Thatcher", "Haiti earthquake", "Philip Seymour Hoffman", "smartphone reviews", "icecream recipes" and "first moon landing") were issued in July 2014 and are meant to provide the readers with a number of different temporal scenarios. While we do depend on Bing's search engine results and dates to have this demo online, our system is not limited to focus on one particular time period and it will tackle different dates as long as they are initially retrieved. With this in mind, we choose both temporal ambiguous and unambiguous queries, as well as queries whose intents are likely to be atemporal or of a more recent temporal nature, in order to span a different number of cases. All the queries except one ("haiti earthquake") are external to the experimental dataset so as to understand the generalization of the results.

Table 10
Top-10 GTE-R (left hand side) and Bing search engine (right hand side) results for the query “haiti earthquake”.

1	**FAST FACTS: Haiti Earthquake. Fox News** Fast facts – Haiti Earthquake USGS: USGS called it the strongest earthquake since 1770 in what is now Haiti The quake struck on January 12, 2010 http://bit.ly/BingSE-40	**Four years after Earthquake, Many in Haiti remain** Four years ago Sunday, a 7.0-magnitude earthquake hit Haiti, destroying its capital of Port-au-Prince and killing more than 200,000 people. Today, much of http://bit.ly/GTERank-23
2	**Haiti Earthquake Relief. The White House** On January 12, 2010, a massive earthquake struck the nation of Haiti, causing catastrophic damage inside and around the capital city of Port-au-Prince. http://bit.ly/BingSE-36	**List of earthquakes in Haiti – Wikipedia, the free** This is a list of earthquakes in Haiti. Some of them have been very destructive to the country. Contents 1 List of major earthquakes 2 12 January 2010 earthquake 3 http://bit.ly/GTERank-17
3	**Haiti Earthquake, Earthquake in Haiti, location map of** Haiti Earthquake, 7.0 Mw Earthquake hits Haiti on January 12, 2010. Earthquake has caused widespread damage near capital city Port-au-Prince and has left over 100000 http://bit.ly/BingSE-15	**Haiti earthquake: where is US aid money going? Get the** American companies and NGOs continue to receive the lion’s share of US aid funding for projects in Haiti four years after the earthquake that levelled the capital http://bit.ly/GTERank-29
4	**Haiti Earthquake Maps – Traveling Haiti .com** At 21:53 UTC on January 12, 2010 an earthquake with a magnitude 7.0 struck the Caribbean nation of Haiti. The US Geological Survey (USGS) says that it was the most http://bit.ly/BingSE-39	**Haiti Earthquake: Pictures, Videos, Breaking News Big News** on Haiti Earthquake. Includes blogs, news, and community conversations about Haiti Earthquake. http://bit.ly/GTE-Rank39
5	**Haiti earthquake – Thomson Reuters Foundation** The 7.0 magnitude quake that rocked Haiti on Jan 12, 2010 was the country’s most powerful in more than 200 years. More than 200,000 people were killed, and 2.3 http://bit.ly/BingSE-16	**Haiti News – Breaking World Haiti News – The New York Times** World news about Haiti. Breaking news and archival information about its people, politics and economy from The New York Times. http://bit.ly/GTERank-44
6	**What are facts about the haiti earthquake – The Q&A wiki** Here are four facts about Haiti Earthquake on Tuesday the 12th of January 2010: It had a magnitude of 7.0 on the Richter scale. It was the worst earthquake in Haiti http://bit.ly/BingSE-44	**Haiti Earthquake Fast Facts – CNN.com** Here’s what you need to know about the 2010 earthquake in Haiti, which struck January 12, 2010. The earthquake measured 7.0 magnitude on the Richter scale http://bit.ly/GTERank-10
7	**The Haiti Earthquake – TIME – TIME – Breaking News** Photos: Haiti One Year Later. Reconstruction of the regions devastated in the January 12, 2010 earthquake proceeds very slowly - if at all More http://bit.ly/BingSE-10	**The forgotten victims of Haiti’s earthquake – Global** MEARDEE. Haiti isn t forgotten there tons of missions and more every day of the week . I LOVE HAITI http://bit.ly/GTERank-50
8	**Magnitude 7.0 – HAITI REGION – USGS Earthquake Hazard** Provides maps, a summary and detailed information on the magnitude 7.0 earthquake that struck Haiti on January 12, 2010 http://bit.ly/BingSE-22	**Haiti Earthquake recovery 3 years Later: Where has the** three years on, much of the conversation surrounding the Haiti earthquake recovery has centered around charities squandering http://bit.ly/GTERank-50
9	**What caused the devastating Haiti earthquake? – Technology** The earthquake that devastated Haiti Tuesday was the strongest temblor to hit the island nation in more than 200 years. The magnitude 7.0 quake caused http://bit.ly/BingSE-31	**Earthquake Haiti News, Photos and Videos – ABC News** Browse Earthquake Haiti latest news and updates, watch videos and view all photos and more. Join the discussion and find more about Earthquake Haiti at abcnews.com http://bit.ly/GTERank-42
10	**Haiti Earthquake Fast Facts – CNN.com** Here’s what you need to know about the 2010 earthquake in Haiti, which struck January 12, 2010. The earthquake measured 7.0 magnitude on the Richter scale http://bit.ly/BingSE-6	**The Haiti Earthquake – TIME – TIME – Breaking news** photos: haiti one year later. Reconstruction of the regions devastated in the January 12, 2010 earthquake proceeds very slowly – if at all More http://bit.ly/GTERank-7

Our first example is provided in Fig. 14 for the query “Margaret Thatcher” to enlighten the readers understanding in relation to the system’s ability in tackling different time periods of the query. By looking at the results, one is able to identify a diversity of different dates related to the query, with particular emphasis on Margaret Thatcher’s birthdate (1925), prime-minister term (1979–1990) and date of her death (2013). It is also worthy to refer the ability of the system in including in the top-5 results a reference to the 1980s when she came to be known as the “Iron Lady”, as well as the inclusion of a text which despite not having any dates, yet is able to convey relevant information, i.e. the age at which she died (87) and the fact that she was the first woman ever to serve as prime minister of Great Britain and the longest-serving British prime minister of the 20th century. Finally, we should call attention to the fact that our system is able to put on the first position of the list of the results a wealth of information on both topic and temporal dimensions. This turns out to be even more relevant as this result was only retrieved by Bing search engine on position number 40. In contrast, Bing’s first result is limited to two dates, i.e., two less than our system for a very similar text.

In our second example (see Table 10) we provide a thorough comparison between the results of our system and Bing’s search engine for the top-10 results query “haiti earthquake”. The URLs of the results were generated using bit.ly and provide position information of its counterpart. For instance, http://bit.ly/BingSE-40, means this GTE-R result appears on position #40 of Bing search engine, whereas, http://bit.ly/GTERank-23, means this Bing search result appears on position #23 of GTE-R. An overall analysis of the results provides the user with a detailed account of the 2010 earthquake. More interestingly however, is the fact that information concerning the 1770 Haiti earthquake can be accessed on GTE-R top-10 results, but not on Bing’s search engine. Another interesting point to mention is that most of the results of Bing search engine for this query remain news-related (e.g., breaking news) though the earthquake occurred four years ago. A few others are not related with the earthquake. This is the case of results number #5 and #7 which are listed by GTE-R in position numbers #44 and #50 respectively.

Next, in Table 11, we show information about “Philip Seymour Hoffman” a famous actor died in 2014. The results listed in the table show the potential of our approach in collecting information for a number of different temporal and topical instances. Indeed, GTE-R gives us additional information about the actor’s cause of death (“overdose”), age at death (“46”) plus the city where he died (“New York”). By looking at result #2 of both lists we can also notice that, though the two texts report to the same

Table 11
Top-5 GTE-R (left hand side) and Bing search engine (right hand side) results for the query “Philip Seymour Hoffman”.

1	**Philip Seymour Hoffman – Wikipedia, the free encyclopedia** Philip Seymour Hoffman (July 23, 1967 – February 2, 2014) was an American actor and director. He was prolific in both film and theater from the early 1990s until http://bit.ly/Bing-SE-1	**Philip Seymour Hoffman – Wikipedia, the free encyclopedia** Philip Seymour Hoffman (July 23, 1967 – February 2, 2014) was an American actor and director. He was prolific in both film and theater from the early 1990s until http://bit.ly/Bing-GTE-Rank-1
2	**Philip Seymour Hoffman News, Pictures, and Videos. TMZ.com** Powered by imdb. Film and stage actor and theater director Philip Seymour Hoffman was born in the Rochester, New York, suburb of Fairport on July 23, 1967. http://bit.ly/Bing-SE-28	**Philip Seymour Hoffman – IMDb** Philip Seymour Hoffman, Actor: Capote. Film and stage actor and theater director Philip Seymour Hoffman was born in the Rochester, New York, suburb of Fairport on http://bit.ly/Bing-GTE-Rank-4
3	**Philip Seymour Hoffman Dead: Actor Dies At 46 In New York** Philip Seymour Hoffman was found dead in his Manhattan apartment, the Wall Street Journal reported Sunday (Feb. 2). The 46-year-old actor’s cause of death http://bit.ly/Bing-SE-35	**Philip Seymour Hoffman – Rotten Tomatoes: Movies. TV** Philip Seymour Hoffman Celebrity Profile - Check out the latest Philip Seymour Hoffman photo gallery, biography, pics, pictures, interviews, news, forums and blogs at http://bit.ly/GTE-Rank-18
4	**Philip Seymour Hoffman – IMDb** Philip Seymour Hoffman, Actor: Capote. Film and stage actor and theater director Philip Seymour Hoffman was born in the Rochester, New York, suburb of Fairport on http://bit.ly/Bing-SE-2	**Philip Seymour Hoffman Biography – Facts, Birthday, Life** Learn more about Philip Seymour Hoffman’s astounding performances as an actor and director, winning him awards and acclaim throughout his career, at Biography.com. http://bit.ly/GTE-Rank-33
5	**Sources: Philip Seymour Hoffman dead of apparent drug** Oscar-winning actor Philip Seymour Hoffman was found dead in his New York home of an apparent drug overdose, law enforcement sources told CNN http://bit.ly/Bing-SE-11	**Philip Seymour Hoffman – ’I Know I’m Gonna Die’. TMZ.com** Philip Seymour Hoffman was on a heroin binge 6 weeks before he died, and told friends he feared he was destined to fatally OD ... TMZ has learned. http://bit.ly/GTE-Rank-42

Table 12
Top-5 GTE-R results for the query “smartphone reviews” (left hand side) and the query “icecream recipes” (right hand side)

1	**Smartphone Review. New Smartphones – Best Smartphones** Latest Smartphone News on New and Upcoming Smartphones Everyone is looking for new smartphone releases in 2013 and 2014, on this site you can find latest news and	**Ice Cream Recipes – Homemade Ice Cream Recipes – Frozen** These ice cream recipes include easy homemade ice cream recipes, frozen yogurt recipes, ice cream cake recipes, ice cream pie recipes and more homemade frozen desserts.
2	**2014 Smartphones – New Smartphones – 2014 New Smartphones** News on latest, upcoming and new smartphones releases in 2012, 2013 and 2014. Daily updates on new smartphones, smartphone reviews, smartphone releases, android	**Easy homemade vanilla ice cream recipe – Allrecipes.com** Use this easy recipe to make vanilla ice cream, or add your favorite flavors to it.
3	**Review Centre: Smartphone Reviews of 2013 and 2014 Mobile** Find out what other buyers really think of the latest smartphones and compare their reviews before you get signed up to a long contract	**Ice cream recipes – Food Network – Easy Recipes, Healthy** Ice Cream Recipes. Treat yourself to frozen favorites from easy homemade ice cream to DIY ice cream sandwiches and more.
4	**Smartphone Reviews – 2014 Phone Reviews and News –** Get expert reviews of 2014 smartphones and find the best smartphone for your needs and read the latest smartphone news, how-to guides and app reviews	**9 Easy homemade ice cream recipes – How to make homemade** making homemade ice cream is easier than you think! Give everybody something to salivate over with these yummy ice cream flavors you can make at home.
5	**BEST SMARTPHONE 2013 – Review and ratings of the best** Smartphone reviews 2013 say that smartphones bring to the user better connectivity than a regular mobile phone because it has a computing platform	**Easy Ice Cream Recipes. Eating Well** Skip store-bought ice cream with ingredients that are hard to pronounce and make your own with our easy homemade ice cream recipes. Whether you’re looking for ...

fact, GTE-R text still gets enriched through the inclusion of an additional temporal reference. The provided evidence, though anecdotal, corroborates our assumption that texts including temporal instances are more informative to the user, as to those not including any temporal reference, thereby contributing to improve the results effectiveness. The results in positions #3 – #5 further confirm that our algorithm is also able to promote to the top, relevant documents which do not include any temporal expression.

In addition, we show two other quite interesting examples, one related to the query “smartphone reviews” for which recent results are likely expectable and another one related to the query “icecream recipes” an atemporal query, for which no temporal results are to be retrieved. Both snapshots (see Table 12), though anecdotal, clearly evidence the ability of our approach in dealing with queries which are far beyond the scope of our research. The “smartphone reviews” query (left hand side of the table) is able to retrieve very recent review results from 2012 onwards, while “icecream recipes” query (right hand side of the table) is capable of retrieving very descriptive topical results.

Finally, we show the tail 5 ranking results (i.e., positions 46–50 of the list of results) for the queries “first moon landing” and “haiti earthquake”. It is interesting to stress that our algorithm is able to position well down in the list of the results, temporally non-relevant documents that were initially listed at top positions by Bing search engine (red color in the table). An overall analysis of the table shows that the first result of Bing for the query “first moon landing (left hand side of Table 13), which is listed by GTE-R in position #46, is the Wikipedia page with a single reference to “moon landing”. The results of this snapshot further confirm that the fact that a snippet may contain a temporal expression is not sufficient to have it classified as relevant to the query. This is particularly evident in GTE-R position #49 (“Page Last Updated: June 27th…”) and confirms our principle that snippets only get promoted to the top list of results, when their contents are temporally but also topically relevant to the query. This is further confirmed for the query “haiti earthquake” (right hand side of Table 13) as GTE-R result #47 is considered to be a non-relevant snippet, yet it includes a temporal expression. More interestingly however are the results listed in GTE-R position

Table 13
Tail-5 GTE-R results for the query “first moon landing” (left hand side) and the query “haiti earthquake” (right hand side)

46	1	**Moon landing – Wikipedia,** the free encyclopedia A moon landing is the arrival of a spacecraft on the surface of the Moon. This includes both manned and unmanned (robotic) missions. The first human-made object to ...	4	**Haiti News – Breaking World Haiti News – The New York Times** World news about Haiti. Breaking news and archival information about its people, politics and economy from The New York Times
47	35	**The First Lunar Landing – NASA** The First Lunar Landing TABLE OF CONTENTS. Introduction Part I Part II Part III Part IV Part V Part VI End	12	**Haiti’s Earthquake, Still Waiting for a…** Carrefour, HAITI, Jan 20 2014 (IPS) –Mimose Gérard sits in her tent at Gaston Margron camp, surrounded by large bags filled with plastic bottles. She
48	20	**The Great Moon Hoax – NASA Science** Fortunately the Soviets didn’t think of the gag first. They could have filmed their own fake Moon landings and really embarrassed the free world.	13	**Haiti Earthquake – CNN** Maxi journeyed to northern Haiti to see Joseph, the only person who could understand her ordeal. They were not friends previously, but after surviving the earthquake ..
49	29	**Apollo 11. NASA** Page Last Updated: June 27th, 2014 Page Editor: NASA Administrator	45	**Haiti Earthquake – YouTube** Molly reports on the recent earthquake in Haiti. Maps of earthquake: ...
50	38	**The First Moon Landing (First Facts: Solar System** The First Moon Landing (First Facts: Solar System) [Kortenkamp, Steve] on Amazon.com. ∗FREE∗ shipping on qualifying offers. Did you know that it took three days for …	5	**The forgotten victims of Haiti’s earthquake – Global** MEARDEE. Haiti isn t forgotten there tons of missions and more every day of the week. I LOVE HAIT

#46 and #50. Both texts clearly evidence that they are not related or relevant to the query, and yet Bing search engine ranked them in position #4 and #5 respectively.

The results presented here show the ability of our system in dealing with different types of queries. As noted previously, our algorithm is particularly targeted to tackle temporal ambiguous and unambiguous queries, i.e., queries having at least one well-defined temporal instance. We have shown however, that atemporal or queries with a more recency nature can also be issued in the search interface, and yet the quality of the results remains intact.

9. Conclusions and perspectives

In this paper, we proposed to adjust the score of a document in a ranking task in response to a given time-sensitive query by following a content-based approach that extracts temporal features from the contents of the document. Our aim was to retrieve, in the top list of results, documents that are not only topically relevant but that are also from the most important time periods, thus contributing to improve results’ effectiveness across a different number of temporal queries. This is a very challenging issue since we need not only to return the most relevant documents that meet the users’ query intents, but also to simultaneously devalue those incorporating non-relevant concepts or dates. For this purpose, we developed GTE-R, a re-ranking model that combines both topical and temporal relevance in a single score. Through extensive experiments, including a crowdsourcing experiment, we demonstrated that GTE-R is able to achieve better results under several evaluation metrics compared to a number of different baselines, including temporal ones. More specifically, we showed that the introduction of the GTE-Class causes an improvement of the GTE-R performance, both in the Top and in the Tail approaches. Moreover, we also showed the behavior of GTE-R under two different types of collections: exclusively temporal ones and a combination of both temporal and atemporal texts. Even though GTE-R performs better under exclusively temporal collections, its effectiveness still gets significantly improved with respect to the baselines when atemporal texts are also considered. Notwithstanding having achieved a good performance, GTE-R is still limited to determine the relevance of a candidate date only in the query context. This can be overcome in the future by enabling GTE-Class to determine the relevance of a candidate date in the context of a document too. As a practical demonstration of our research, we also provide GTE-R as web search interface to the research community. Additionally, and as a further contribution in the context of temporally re-ranking web snippets within time-sensitive queries, we also made available two gold standard datasets that set baselines for future research. Although we focused on web snippets in our experiments, our approach might similarly be applicable to collections of short texts embodying temporal information, such as Twitter posts.

Acknowledgments

The authors want to thank the participants of the user study. This work was partially funded by Project "NORTE-07-0124-FEDER-000059" which is financed by the North Portugal Regional Operational Programme (ON.2 – O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT) within Project UID/EEA/50014/2013. It was also financed by the Center of Mathematics, University of Beira Interior, within Project «PEst-OE/MAT/UI0212/2014».

References

Alonso, O., Baeza-Yates, R., & Gertz, M. (2009). Effectiveness of temporal snippets. In *Proceedings of the workshop on web search result summarization and presentation (WSSP) associated to the 18th international world wide web conference (WWW)*. ACM Press April 20–24.

Alonso, O., Gertz, M., & Baeza-Yates, R. (2011). Enhancing document snippets using temporal information. In Roberto Grossi, et al. (Eds.), *Proceedings of the18th international symposium on string processing and information retrieval (SPIRE) (Lecture notes in computer science)* (pp. 26–31). Berlin/Heidelberg, Pisa, Italy: Springer October 17–21.

Amati, G. (2003). *Probabilistic models for information retrieval based on divergence from randomness*. Scotland, UK: School of Computing Science, University of Glasgow (Ph.D. thesis).

Berberich, K., Vazirgiannis, M., & Weikum, G. (2005). Time-aware authority ranking. *Internet Mathematics, 2*(3), 301–332.

Berberich, K., Bedathur, S., Alonso, O., & Weikum, G. (2010). A language modeling approach for temporal information needs. In *Proceedings of the 32nd European conference on information retrieval (ECIR) (Lecture notes in computer science – research and advanced technology for digital libraries)* (pp. 13–25). Springer-Verlag. March 28–31.

Berberich, K., & Bedathur, S. (2013). Temporal diversification of search results. In *Proceedings of the workshop on time-aware information access (TAIA) associated to the 36th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* August 1.

Callan, J., & Moffat, A. (2012). Panel on use of proprietary data. *ACM SIGIR Forum, 46*(2), 10–18.

Campos, R., Dias, G., & Jorge, A. M. (2011). What is the temporal value of web snippets? In *Proceedings of the 1st international temporal web analytics workshop (TWAW) associated to the 20th international world wide web conference (WWW)* (pp. 9–16). March 28: CEUR Workshop Proceedings.

Campos, R., Dias, G., Jorge, A. M., & Nunes, C. (2012). GTE: a distributional second-order co-occurrence approach to improve the identification of top relevant dates. In *Proceedings of the 21st international conference on knowledge and information management (CIKM)* (pp. 2035–2039). ACM Press. October 29–November 02.

Campos, R., Dias, G., Jorge, A. M., & Nunes, C. (2014a). GTE-Cluster: a temporal search interface for implicit temporal queries. In *Proceedings of the 36th European conference on information retrieval (ECIR) (Lecture notes in computer science – advances in information retrieval, 8416/2014)* (pp. 775–779). Springer-Verlag. April 13–16.

Campos, R., Dias, G., Jorge, A. M., & Jatowt, A. (2014b). Survey of temporal information retrieval and related applications. *ACM Computing Surveys, 47*(2), 1–41 Article 15.

Campos, R., Dias, G., Jorge, A. M., & Nunes, C. (2014c). GTE-Rank: searching for implicit temporal query results. In *Proceedings of 23rd ACM international conference on information and knowledge management (CIKM)* (p. 2081). ACM Press. November 03–07.

Chang, A., & Manning, C. (2012). SUTIME: a library for recognizing and normalizing time expressions. In *Proceedings of the 8th international conference on language resources and evaluation (LREC)* May 23–25.

Chang, P.-T., Huang, Y-C., Yang, C.-L., Lin, S-D., & Cheng, P-J. (2012). Learning-based time-sensitive re-ranking for web search. In *Proceedings of the 35th annual international ACM conference on research and development in information retrieval (SIGIR)* (pp. 1101–1102). ACM Press. August 12–16.

Cheng, S., Arvanitis, A., & Hristidis, V. (2013). How fresh do you want your search results? In *Proceedings of the 22nd international conference on knowledge and information management (CIKM)* (pp. 1271–1280). ACM Press. October 27–November 01.

Church, K. W., & Hanks, P. (1990). Word association norms mutual information and lexicography. *Computational Linguistics, 16*(1), 23–29.

Croft, W. B., Metzler, D., & Strohman, T. (2009). *Search engines: information retrieval in practice*. Addison Wesley.

Dai, N., Shokouhi, M., & Davison, B. D. (2011). Learning to rank for freshness and relevance. In *Proceedings of the 34th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 95–104). ACM Press. July 24–28.

Dakka, W., Gravano, L., & Ipeirotis, P. G. (2012). Answering general time sensitive queries. *IEEE Transactions on Knowledge and Data Engineering, 24*(2), 220–235 IEEE Computer Society Press.

Dias, G., Alves, E., & Lopes, J. (2007). Topic segmentation algorithms for text summarization and passage retrieval: an exhaustive evaluation. In *Proceedings of the 22th conference on artificial intelligence (AAAI)* (pp. 1334–1340). AAAI Press. July 22–26.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecological Society of America, 26*, 297–302.

Dong, A., Chang, Y., Zheng, Z., Mishne, G., Bai, J., Zhang, R., et al. (2010). Towards recency ranking in web search. In *Proceedings of the 3rd ACM international conference on web search and data mining (WSDM)* (pp. 11–20). ACM Press. February 3–6.

Efron, M., & Golovchinsky, G. (2011). Estimation methods for ranking recent information. In *Proceedings of the 34th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 495–504). ACM Press. July 24–28.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382.

Gey, F., Larson, R., Machado, J., & Yoshioka, M. (2011). NTCIR9-GeoTime overview – evaluating geographic and temporal search: round 2. In *Proceedings of the 9th NTCIR workshop (NTCIR-9)* (pp. 9–17). December 6–9.

Gey, F., Larson, R., Kando, N., Machado, J., & Sakai, T. (2010). NTCIR-GeoTime overview: evaluating geographic and temporal search. In *Proceedings of the 8th NTCIR workshop (NTCIR-8)* (pp. 147–153). June 15–18.

Guo, Q., Diaz, F., & Yom-Tov, E. (2013). Updating users about time critical events. *Advances in Information Retrieval (Lecture Notes in Computer Science), 7814*, 483–494.

Hiemstra, D. (2001). *Using language models for information retrieval*. Netherlands: Centre for Telematics and Information Technology, University of Twente (Ph.D. thesis).

Joho, H., Jatowt, A., & Blanco, R. (2014). NTCIR temporalia: a test collection for temporal information access research. In *Proceedings of the 4th temporal web analytics workshop (TempWeb4) associated to the 23rd international world wide web conference (WWW)* (pp. 845–849). International World Wide Web Conferences Steering Committee. April 8.

Jones, R., & Diaz, F. (2007). Temporal profiles of queries. *ACM Transactions on Information Systems, 25*(3) Article No. 14.

Kanhabua, N., & Nørvåg, K. (2010). Determining time of queries for re-ranking search results. In *Proceedings of 14th European conference on digital libraries (ECDL)* (pp. 261–272). September 6–10.

Kanhabua, N., & Nørvåg, K. (2012). Learning to rank search results for time-sensitive queries. In *Proceedings of the 21st international conference on knowledge and information management (CIKM)* (pp. 2463–2466). ACM Press. October 29–November 02.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika, 30*(1–2), 81–93.

Kumar, R., & Vassilvitskii, S. (2010). Generalized distances between rankings. In *Proceedings of the 19th international world wide web conference (WWW)* (pp. 571–579). ACM Press. April 26–30.

Li, X., & Croft, W. B. (2003). Time-based language models. In *Proceedings of the 12th international conference on knowledge and information management (CIKM)* (pp. 469–475). ACM Press. November 2–8.

Machado, D., Barbosa, T., Pais, S., Martins, B., & Dias, G. (2009). Universal mobile information retrieval. In *Proceedings of the 13th international conference on human–computer interaction (HCII)* (pp. 345–354). July 19–24.

Metzler, D., Jones, R., Peng, F., & Zhang, R. (2009). Improving search relevance for implicitly temporal queries. In *Proceedings of the 32nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 700–701). ACM Press.

Nunes, S., Ribeiro, C., & David, G. (2007). Using neighbors to date web documents. In *Proceedings of the 9th ACM international workshop on web information and data management (WIDM) associated to the 16th international conference on knowledge and information management (CIKM)* (pp. 129–136). ACM Press. November 9.

Robertson, S. E., Stephen, E., Walker, S., Jones, S., Hancock-Beaulieu, M., & Gatford, M. (1994). Okapi at TREC-3. In *Proceedings of the third text retrieval conference (TREC)* (pp. 109–126).

Scaiella, U., Ferragina, P., Marino, A., & Ciaramita, M. (2012). Topical clustering of search results. In *Proceedings of the 5th ACM international conference on web search and data mining (WSDM)* (pp. 223–232). ACM Press. February 8–12.

Silva, J. F., Dias, G., Guilloré, S., & Pereira, J. G. (1999). Using LocalMaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. In *Proceedings of the 9th Portuguese conference in artificial (EPIA)* (pp. 21–24). September 21–24.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11–21.

Spärck Jones, K., & Van Rijsbergen, C. J. K. (1975). *Report on the need for and provision of an 'Ideal' information retrieval test collection: In British library research and development report 5266*. Cambridge: University Computer Laboratory.

Spärck Jones, K., & Bates, R. G. (1977). *Report on a design study for the "Ideal" information retrievaltest collection: In British library research and development report 5488*. Cambridge: University Computer Laboratory.

Spärck Jones, K., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Information Processing and Management: An International Journal, 36*(6), 779–840.

Spearman, C. (1987). The proof and measurement of association between two things. By C. Spearman, 1904. *The American Journal of Psychology, 100*(3-4), 441–471 (r 1987):.

Strötgen, J., & Gertz, M. (2010). HeidelTime: high quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th international workshop on semantic evaluation (IWSE) associated to the 41th annual meeting of the association for computational linguistics (ACL)* (pp. 321–324). July 11–16.

Styskin, A., Romanenko, F., Vorobyev, F., & Serdyukov, P. (2011). Recency ranking by diversification of result set. In *Proceedings of 20th international conference on knowledge and information management (CIKM)* (pp. 1949–1952). ACM Press. October 24–28.

Zhang, R., Chang, Y., Zheng, Z., Metzler, D., & Nie, J-Y. (2009). Search result re-ranking by feedback control adjustment for time-sensitive query. In *Proceedings of the North American chapter of the association for computational linguistics – human language technologies (NAACL)* (pp. 165–168). May 31–June 5.

Zobek, J. (1998). How reliable are the results of large-scale retrieval experiments? In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 307–314). ACM Press. August 24–28.