GTE-Rank: Searching for Implicit Temporal Query Results

Ricardo Campos, Gaël Dias, Alipio Jorge, Celia Nunes

GTE-Rank: Searching for Implicit Temporal Query Results

Ricardo Campos [1,2,6], Gaël Dias [4], Alípio Mário Jorge [2,3], Célia Nunes [5,6]

[1] Polytechnic Institute of Tomar, Portugal; [2] LIAAD – INESC TEC; [3] DCC – FCUP, University of Porto, Portugal; [4] HULTECH/GREYC, University of Caen Basse-Normandie, France; [5] Department of Mathematics, University of Beira Interior, Covilhã, Portugal; [6] Center of Mathematics, University of Beira Interior, Covilhã, Portugal

ricardo.campos@inesctec.pt, gael.dias@unicaen.fr, amjorge@fc.up.pt, celian@ubi.pt

ABSTRACT

Temporal information retrieval has been a topic of great interest in recent years. Despite the efforts that have been conducted so far, most popular search engines remain underdeveloped when it comes to explicitly considering the use of temporal information in their search process. In this paper we present GTE-Rank, an online searching tool that takes time into account when ranking time-sensitive query web search results. GTE-Rank is defined as a linear combination of topical and temporal scores to reflect the relevance of any web page both in topical and temporal dimensions. The resulting system can be explored graphically through a search interface made available for research purposes.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Query Formulation*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *Performance*

Keywords

Temporal Information Retrieval, Implicit Temporal Queries, Temporal Re-Ranking, Temporal Query Understanding.

1. INTRODUCTION

Despite the emergence of a large spectrum of time-aware search and retrieval applications, most popular search engines still do not take advantage of the temporal information contained in web pages. Current search engines 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 web search results. While this may be a suitable solution for the news domain for which a huge 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 wide coverage temporally diversified information. This is particular evident for implicit temporal queries (e.g., "*Haiti earthquake*", "*BP oil spill*" or "*Madagascar*"), which beyond not being explicitly tagged with a temporal feature, span over a broad timeline. In this paper, we present GTE-Rank an online search interface, which allows searching for topics through time. GTE-Rank is designed to enhance user's experience through a balanced approach that takes into account both the conceptual and the temporal dimensions of the topic. 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 contextual perspective. For instance, when querying for the well-known American actor "*Philip Seymour Hoffman*", who has passed away recently, it would be interesting to know who he was or when did he die, but also to be provided with other important topics and time information, such as, when did he begin acting in television, where he was born or which movie gave him the award for the best actor performance. Such a new presentation of the results would enable users to gain not only a broad insight of the topic but also to understand its multiple temporal dimensions, thus contributing to improve user's satisfaction [1].

Aware of the above, 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. Different studies have been proposed to solve this problem. The research that are most related to our approach are [2,7,8]. Berberich *et al*. [2] for example, ranks documents according to the estimated probability of generating the query through a language model framework that requires documents and queries to be explicitly time stamped. The methods put forward by Metzler *et al*. [7] and Kanhabua & Nørvåg [8] suggest an alternative solution. They propose a time-dependent ranking model to explicitly adjust the score of a document in favor of those matching the determined time(s) of an implicit temporal query. None of these works however, consider extracting temporal features from the documents contents in order to determine the possible time(s) of the query. Metzler *et al*. [7] uses query logs and Kanhabua & Nørvåg [8] takes advantage of the creation date of the document, which may be significantly different from the actual content. We differ from previous studies on this subject in several other aspects. First, our methodology is unsupervised as no specific training process is needed to determine the time(s) of the query. Second, it is mostly language-independent as it implements a rule-based model supported by simple regular expressions to extract relevant dates. 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. The remainder of this paper is organized as follows. Section 2 describes the architecture behind the GTE-Rank system. Section 3 presents the proposed demonstration. Finally, Section 4 concludes this paper with some final remarks.

2. SYSTEM ARCHITECTURE

In this section, we describe the main components of the GTE-Rank system. GTE-Rank proceeds in two steps. First, the system evaluates the correlation between a given query and the candidate dates extracted within web snippets leading to the identification of top relevant dates. Second, a linear combination of topical and temporal scores is defined to reflect the relevance of any web snippet both in the topical and in the temporal dimensions. The overall structure of GTE-Rank architecture is represented in Figure 1 and consists of five different modules: (1) Web search; (2) Web snippet representation; (3) Temporal similarity; (4) Date filtering and (5) Temporal ranking.

CIKM '14 , Nov 03-07 2014, Shanghai, China
ACM 978-1-4503-2598-1/14/11.
http://dx.doi.org/10.1145/2661829.2661856

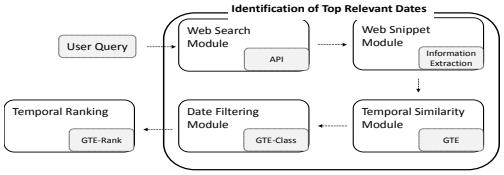

Figure 1: GTE-Rank Architecture.

The GTE-Rank interface receives a query from the user, fetches related web snippets from a given search engine and applies text processing to all web snippets. This processing task involves selecting the most relevant words and collecting the candidate years in each web snippet. Each candidate year is then given a temporal similarity value to the query computed in the temporal similarity module. We then apply a classification strategy in the date filtering module, to determine whether the candidate years are actually relevant or not to the query. Non-relevant ones will be simply discarded by the system. Each snippet is then reordered according to its contents and how they relate to the query both in the temporal and in the topical dimensions. A brief description of each component is provided below.

(1) Web Search. We apply a web search API, which, given an implicit temporal query, accesses an up-to-date index search engine to obtain a collection of web results. 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.

(2) Web Snippet Representation. Each snippet is represented by a bag-of-relevant-words and a set of candidate temporal expressions extracted from the title of the snippet and from the text itself. As shown by Alonso *et al*. [1], web snippets offer an interesting alternative for the representation of web documents, where years often appear, thus avoiding the cost of parsing full web pages. We rely on a segmentation process and a numerical selection heuristic to extract relevant words[1] and a simple rule-based model supported on regular expressions to extract explicit temporal patterns. We focus on the extraction of temporal patterns of the year granularity level to keep the system mostly language-independent. The obtained result is a set of distinct candidate years extracted from the set of all web snippets.

(3) Temporal Similarity. Each candidate year d_j is then given a temporal similarity value representing its degree of relevance to the query q. To model this relevance, we apply our temporal similarity measure GTE [3], which retrieves a value ranging from 0 to 1. A web service of GTE is provided[2] 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.

(4) Date Filtering. Next, the system determines whether or not the candidate temporal expressions are relevant to the query by applying GTE-Class [3], a classical threshold-based strategy, which considers a candidate date to be relevant, if and only if $GTE(q, d_j) \geq \lambda$ and non-relevant otherwise. Based on this, each snippet is no longer represented by a set of candidate temporal expressions but by a set of relevant dates. One consequence of this, is a direct impact on the quality of the retrieved results, as non-relevant or wrong dates are simply discarded. A description of the GTE-Class demo[3] can be found in our recent work [5].

(5) Temporal Ranking. The final step of the GTE-Rank architecture is our temporal re-ranking model. GTE-Rank relies on a linear combination approach that considers topical and temporal scores. The underlying idea is that a document should be ranked higher if its contents are topically and temporally related to the query. GTE-Rank is defined below. $\alpha \in [0,1]$ and $\beta = 1 - \alpha$ are the tuning parameters 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, $w_{h,i} \in W_{S_i}, h = 1,..,k$ is one of the k most relevant terms of the snippet S_i and IS [6] a second-order similarity measure that calculates the correlation between all pairs of two context vectors X and Y, where X is the context vector representation of q and Y of $w_{h,i}$. Both context vectors are formed by a combination of the best relevant terms and best relevant dates determined by the DICE coefficient measure.

$$GTE\text{-}Rank(q, S_i) = \alpha \sum_{j=1}^{u} GTE\left(q, d_{j,i}^{Rel}\right) + \beta \sum_{h=1}^{k} IS(q, w_{h,i})$$

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 snippet. IS gives the similarity between the query and each of the relevant words found in the snippet. Note that one of the advantages of our approach relies precisely on the use of GTE as it enables GTE-Class to filter out from the ranking module the set of all non-relevant dates. Experiments with a publicly available dataset[4] consisting of 1900 web snippets and 38 implicit text queries show that GTE-Rank is able to achieve better results under several evaluation metrics compared to three different baselines. A fully detailed description of the underlying scientific approach and the evaluation methodology can be found in [4].

3. DEMONSTRATION OVERVIEW

As a result of our research, we publicly provide an online demo (http://wia.info.unicaen.fr/GTERankAspNet_Server).GTE-Rank was implemented using .Net technology (C#) and asp.net on the server side. The implemented version is designed to demonstrate the current state of the demo, thus concerns of design nature where not taken into account. Although the main motivation of our work is focused on queries with temporal nature, GTE-Rank allows the execution of any query including non-temporal ones. Since our system does not pose any constraint in terms of language or domain, users can issue queries in any language, ranging from business (e.g. "*iPad*"), cinema (e.g. "*true grit*"), politics (e.g. "*Margaret Thatcher*"), natural disasters (e.g. "*Haiti earthquake*"), musical topics (e.g. "*Radiohead*"), to cite just a few. To retrieve the results, we use a prospective search where the query is first issued before results are gathered and indexed. For this purpose we rely on Bing Search API[5] with the *en-US* language parameter defined to retrieve 50 results per query. The proposed solution is computationally efficient and can easily be tested online (limited to 5000 queries per month). In response to a query submitted in a search box, GTE-Rank 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 conceptual 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

[1] http://wia.info.unicaen.fr/TokenExtractor/api/Token?query= [May 29th, 2014]

[2] http://wia.info.unicaen.fr/GTEAspNetFlatTempCluster_Server/api/GTE?FilterDates=false&query=

[3] http://wia.info.unicaen.fr/GTEAspNetFlatTempCluster_Server [May 29th, 2014]

[4] http://www.ccc.ipt.pt/~ricardo/datasets/WCRank_DS.html [May 29th, 2014]

[5] https://datamarket.azure.com/dataset/5BA839F1-12CE-4CCE-BF57-A49D98D29A44 [May 29th, 2014]

preset to 0.8 as GTE-Rank 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. In this demo, we show the ability of the ranking system not only in how it pulls up to the top the relevant documents, but also in how it pushes down to the tail the non-relevant ones, thus ensuring that they will not occupy top positions of the ranking results. An illustration of the interface is provided in Figure 2 for the query "*true grit*" (top 10 results). It is interesting to note that our algorithm retrieves in the second, third, sixth, seventh and tenth position, five relevant results that were initially retrieved by the Bing search engine in the thirty-fifth, thirty-first, twenty-first, thirty-ninth and twenty-fifth positions, respectively. Furthermore, we show that our algorithm is also able to promote to the top, relevant documents which do not include any temporal expression.

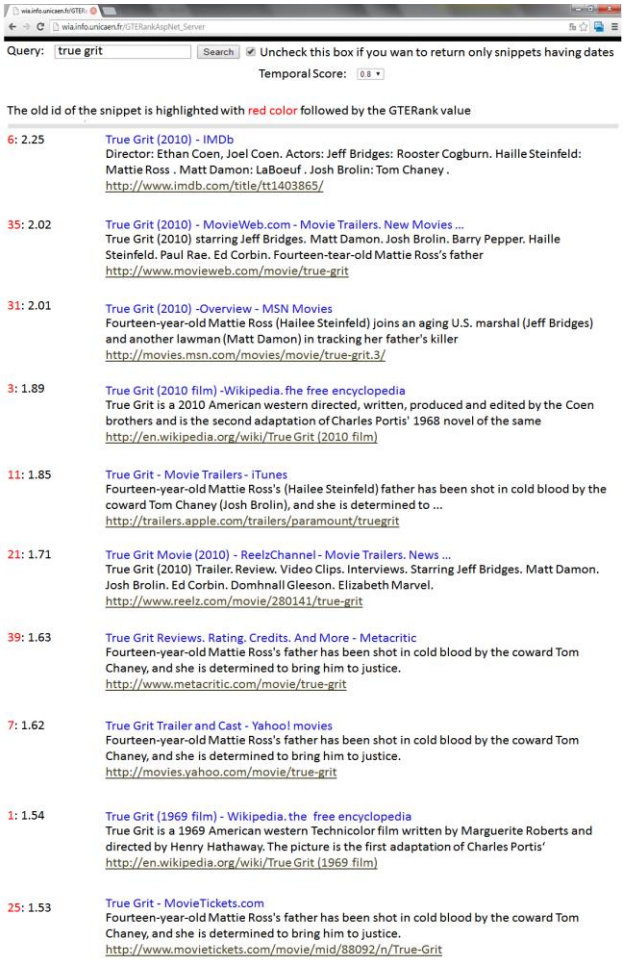

Figure 2: GTE-Rank interface for the query "true grit". Top 10.

Finally, Figure 3 shows the tail 5 ranking results for the same query. It is interesting to note that our algorithm is able to position well down in the list of the results, temporally non-relevant documents that were initially positioned at top positions by Bing search engine, of which IDs 5, 13 and 17 are elucidative examples. A video outlining the demo proceeding is available at http://www.ccc.ipt.pt/~ricardo/software.html.

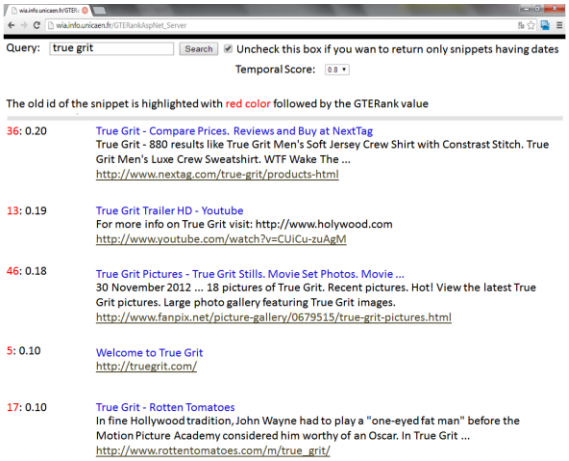

Figure 3: GTE-Rank interface for the "true grit". Tail 5.

4. CONCLUSION

In this article we present GTE-Rank - an online searching system that aims to retrieve in the top list of the results, documents that are not only topically relevant but that are also from the most important time periods. GTE-Rank relies on a similarity measure that is capable of identifying top relevant dates for queries where no temporal information is provided and a re-ranking model that combines both conceptual and temporal relevancies in a single score, thus offering a balanced approach of the results. We adopt a methodology that can be applied to real-world search scenarios and a content-based approach, which enables to return documents about a given period, as opposed to the retrieval of documents written or published at a given date. As a practical demonstration of our research, we provide a demo service so that GTE-Rank can be tested by the research community. Although we focus on web snippets, our approach is similarly applicable to small texts collections embodying temporal information, such as Twitter posts.

5. ACKNOWLEDGMENTS

This research was 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 FCT. It was also financed by the ERDF through the COMPETE Programme (operational programme for competitiveness), by National Funds through the FCT within project «FCOMP-01-0124-FEDER-037281» and by the Center of Mathematics, UBI, within project «PEst-OE/MAT/UI0212/2014».

6. REFERENCES

[1] Alonso, O., Baeza-Yates, R., and Gertz, M. (2009). Effectiveness of Temporal Snippets. In WSSP'09-WWW'09. Madrid, Spain.

[2] Berberich, K., Bedathur, S., Alonso, O., and Weikum, G. (2010). A Language Modeling Approach for Temporal Information Needs.

[3] Campos, R., Dias, G., Jorge, A. M., and Nunes, C. (2012). GTE: A Distributional Second-Order Co-Occurrence Approach to Improve the Identification of Top Relevant Dates. In CIKM'12.

[4] Campos, R. (2013). Disambiguating Implicit Temporal Queries for Temporal Information Retrieval Applications. PhD Thesis. UP, Portugal.

[5] Campos, R., Dias, G., Jorge, A. M., and Nunes, C. (2014). GTE-Cluster: A Temporal Search Interface for Implicit Temporal Queries. In ECIR'14.

[6] Dias, G., Alves, E., and Lopes, J. (2007). Topic Segmentation Algorithms for Text Summarization and Passage Retrieval: An Exhaustive Evaluation. In AAAI'07. Canada.

[7] Metzler, D., Jones, R., Peng, F., and Zhang, R. (2009). Improving Search Relevance for Implicitly Temporal Queries. In SIGIR'09.

[8] Kanhabua, N., and Nørvåg, K. (2010). Determining Time of Queries for Re-Ranking Search Results. In ECDL'10. Scotland.