GTE-Cluster: A Temporal Search Interface for Implicit Temporal Queries

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

[1]LIAAD – INESC TEC
[2]Polytechnic Institute of Tomar, Portugal
[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@ipt.pt, gael.dias@unicaen.fr, amjorge@fc.up.pt, celian@ubi.pt

Abstract. In this paper, we present GTE-Cluster an online temporal search interface which consistently allows searching for topics in a temporal perspective by clustering relevant temporal Web search results. GTE-Cluster is designed to improve user experience by augmenting *document relevance* with *temporal relevance*. The rationale is that offering the user a comprehensive temporal perspective of a topic is intuitively more informative than retrieving a result that only contains topical information. Our system does not pose any constraint in terms of language or domain, thus users can issue queries in any language ranging from business, cultural, political to musical perspective, to cite just a few. The ability to exploit this information in a temporal manner can be, from a user perspective, potentially useful for several tasks, including user query understanding or temporal clustering.

Keywords: Temporal Information Retrieval, Temporal Clustering, Implicit Temporal Queries

1 Introduction

In recent years, a large number of temporal applications have been developed, mostly concerning Web archives (e.g. Internet Archive [1]), temporal taggers (e.g. HeidelTime [10]), temporal and spatial knowledge bases (e.g. Yago2 [5]), new forms of visualizing temporal data (e.g. SIMILE Timeline Visualization[1]), applications to track how topics evolve over time (e.g. Time Explorer [7], Google nGram viewer [8]), but also commercial services like recordedfuture.com

Despite a clear improvement of search and retrieval temporal related applications, current search engines are still mostly unaware of the temporal dimension. Indeed, in most cases, systems are limited to offer the user the chance to restrict the search to a particular time period or to simply ask him to explicitly specify a time span. If the user is not explicit in his search intents (e.g. "*los angeles earthquakes*") search engines may likely fail to present an overall historic perspective of the topic. Most search engines also provide query auto-completion suggestions to users after they start typing their query in the search box, but usually lack to include suggestions of temporal nature [1]. Similar problems can be observed for query re-formulation suggestions which are shown to the users after they submit their initial query. In both cases, query suggestions rely on past popularity of matching queries thus depending

[1] http://www.simile-widgets.org/timeline/ [October 28th, 2013]

on the user's own knowledge and on the fact that some versions of the query have already been issued. While this works rather well for text suggestions it may cause some problems in case of temporal ones due to the fact that users are usually silent when it comes to explicitly express their temporal intents [9]. They are also largely unaware of the temporal dimension when the query is topically and temporally ambiguous (e.g. "*Madagascar*"). They may be able to detect the different facets of the query (country and movie), but remain alien of the fact that each one may have a different temporal nature.

The examples laid out above show that an end-to-end temporal retrieval system that consistently exploits temporal information is still to be seen. Such a retrieval system would be able to offer the user a temporal overview of the query and to provide information on its various temporal dimensions. But, it should also be able to only present the most relevant dates thus helping to improve the user satisfaction while meeting his information needs. For example, when querying "*margaret thatcher*", it would be interesting to have a few separate clusters (e.g. {1925, 1979, 1990, 2013}) highlighting the most important time periods (birth date, prime-minister period and death date) of this well-known British prime-minister.

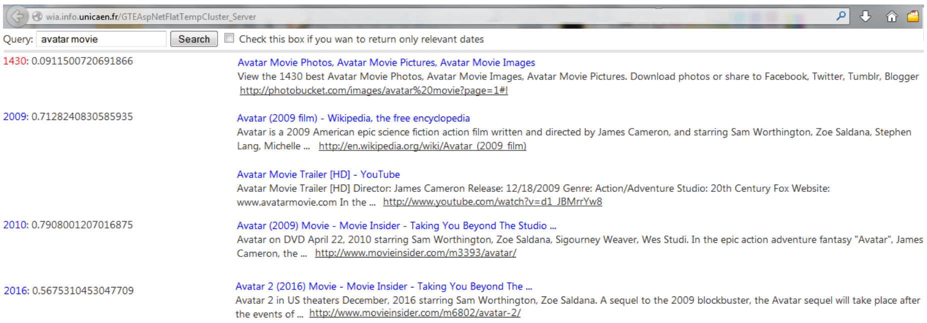
This paper presents the GTE-Cluster temporal search interface which implements a flat temporal clustering model that groups documents at the year level. Our method is based on (1) the identification of relevant temporal expressions extracted from Web snippets and (2) a clustering methodology, where documents are grouped into the same cluster if they share a common year. The resulting clusters directly reflect groups of individual years that show a high connectivity to the text query. One such presentation of the results enables users to have a quick overview of a topic, without the need to go through an extensive list of results. As a result of our research, we publicly provide an online demo, which allows the execution of different kinds of queries, such as business (e.g. "*iPad*"), cultural (e.g. "*avatar movie*"), musical (e.g. "*Radiohead*") or natural disaster ones (e.g. "*Haiti earthquake*"). Although the main motivation of our work is focused on queries with temporal nature, the implemented prototypes allow the execution of any query including non-temporal ones.

2 System Overview

GTE-Cluster consists of five modules: Web search, Web snippet representation, temporal similarity, date filtering and temporal clustering. The demo 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/multiwords and collecting the candidate years in each Web snippet. Each candidate year is then given a temporal similarity value to the query computed by the GTE metric [3] 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 clustered according to its associated years on the assumption that two snippets are temporally similar if they are highly related to the same set of dates. Since any Web snippet can contain several different relevant years, overlapping is allowed. The final set of clusters consists of *m* entities, where *m* is the number of relevant years. The temporally tagged *m* clusters are then sorted in ascending order. One of the advantages of our clustering model is that instead of considering all the temporal expressions as equally relevant as in [1], we determine which ones are more relevant to the user text query. One consequence of this is a direct impact on the quality of the retrieved clusters, as non-relevant or wrong dates are discarded. An evaluation of our approach using several performance metrics, a comparison against a well-known open-source Web snippet clustering engine and a user study demonstrates that GTE-Cluster improves the effectiveness of current approaches. Detailed results of our algorithm are available in [3,4].

3 Demo

The results of our research can be graphically explored by a demo search interface (http://wia.info.unicaen.fr/GTEAspNetFlatTempCluster_Server) made publicly available for research purposes and a video[2]. The implemented version is designed to demonstrate the current state of the demo, thus concerns of design nature where not taken into account. GTE-Cluster is designed to help users searching for information of a given topic through time without any temporal constraint. We rely on Bing Search API[3] with the *en-US* language parameter 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-Cluster displays a set of clusters generated on the fly, which can be instantly used for interactive browsing purposes. We offer two types of retrieval: one that returns only the relevant clusters (marked in blue) and one that combines relevant clusters with non-relevant ones (marked in red). Each cluster is assigned a temporal similarity value reflecting its similarity with the user query. This allows users to not only have an overall perspective of the relevance of the results, but also to evaluate the systems' effectiveness regarding decisions about the relevance of a temporal cluster. An illustration of the interface is provided in Fig. 1 for the query "*avatar movie*".



Query: avatar movie

□ Check this box if you wan to return only relevant dates

1430: 0.0911500720691866

Avatar Movie Photos, Avatar Movie Pictures, Avatar Movie Images
View the 1430 best Avatar Movie Photos, Avatar Movie Images, Avatar Movie Pictures. Download photos or share to Facebook, Twitter, Tumblr, Blogger
http://photobucket.com/images/avatar%20movie?page=1#!

2009: 0.7128240830585935

Avatar (2009 film) - Wikipedia, the free encyclopedia
Avatar is a 2009 American epic science fiction action film written and directed by James Cameron, and starring Sam Worthington, Zoe Saldana, Stephen Lang, Michelle ... http://en.wikipedia.org/wiki/Avatar_(2009_film)

Avatar Movie Trailer [HD] - YouTube
Avatar Movie Trailer [HD] Director: James Cameron Release: 12/18/2009 Genre: Action/Adventure Studio: 20th Century Fox Website: www.avatarmovie.com In the ... http://www.youtube.com/watch?v=d1_JBMrrYw8

2010: 0.7908001207016875

Avatar (2009) Movie - Movie Insider - Taking You Beyond The Studio ...
Avatar on DVD April 22, 2010 starring Sam Worthington, Zoe Saldana, Sigourney Weaver, Wes Studi. In the epic action adventure fantasy "Avatar", James Cameron, the ... http://www.movieinsider.com/m3393/avatar/

2016: 0.5675310453047709

Avatar 2 (2016) Movie - Movie Insider - Taking You Beyond The ...
Avatar 2 in US theaters December, 2016 starring Sam Worthington, Zoe Saldana. A sequel to the 2009 blockbuster, the Avatar sequel will take place after the events of ... http://www.movieinsider.com/m6802/avatar-2/

Fig. 1. GTE-Cluster interface for the query "*avatar movie*".

The values in front of the cluster reflect the similarity value computed by the GTE similarity measure. Note that clusters with a similarity value < 0.35 are considered non-relevant. In contrast, relevant clusters are marked in blue. It is worth noting that our algorithm is capable of detecting as non-relevant the clusters labeled as 1430, while detecting the most relevant ones, i.e., 2009, 2010 and 2016.

[2] http://www.ccc.ipt.pt/~ricardo/software.html [October 28th, 2013]

[3] https://datamarket.azure.com/dataset/5BA839F1-12CE-4CCE-BF57-A49D98D29A44 [October 28th, 2013]

4 Conclusion and Future Work

In this paper, we presented GTE-Cluster a temporal search interface that focuses on disambiguating a text query with respect to its temporal purpose(s). We proposed a strategy for temporal clustering of Web search results, where snippets are clustered by year. We believe that the introduction of the temporal dimension will help to mitigate the limitations that users experience when their information needs include topics of a temporal nature. This is our first version approach to flat temporal clustering of search results. While we already achieved an initial stage of flat clustering by time, our proposal still lacks an approach focused on topics. We are aware that our solution is, from a clustering point of view, a straightforward algorithm. In spite of that, we believe this can open up the debate and create opportunities for future research improvements. As future research, we aim to provide an effective clustering algorithm that clusters and ranks snippets, both based on their temporal and conceptual proximities. A future approach should also consider a more elaborated mechanism in terms of ranking by applying an inter-cluster and an intra-cluster solution. This will enable to reduce the user's effort thus avoiding the need to go through all the clusters and snippets to find the most relevant one. Finally, we may use the similarity value associated to each cluster to offer an ordered set of temporal query suggestions. It would also be useful to consider re-formulations of the initial query for each identified time period as different terms may be associated to different years.

5 Acknowledgments

This work is financed by the ERDF – European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT (Portuguese Foundation for Science and Technology) within project «FCOMP-01-0124-FEDER-037281».

References

1. Alonso, O., Gertz, M., Baeza-Yates, R. Clustering and Exploring Search Results using Timeline Constructions. In: CIKM'09. pp. 97-106. ACM Press, (2009).
2. Campos, R., Jorge, A. Dias, G. Using Web Snippets and Query-logs to Measure Implicit Temporal Intents in Queries. In: QRU'11 associated to SIGIR'11. pp. 13-16. (2011).
3. Campos, R., Dias, G., Jorge, A. M., Nunes, C. GTE: A Distributional Second-Order Co-Occurrence Approach to Improve the Identification of Top Relevant Dates. In: CIKM'12. pp. 2035-2039. (2012).
4. Campos, R., Jorge, A. M., Dias, G., Nunes, C. Disambiguating Implicit Temporal Queries by Clustering Top Relevant Dates in Web Snippets. In: WIC'12. pp. 1-8. IEEE, (2012).
5. Hoffart, J., Suchanek, F.M., Berberich, K., Lewis-Kelham, E., de Melo, G., Weikum, G. YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and many Languages. In: WWW'11. pp. 331-340. ACM Press, (2011).
6. Kahle, B. Preserving the Internet. Scientific American Magazine, 276(3). pp. 72-73. (1997).
7. Matthews, M., Tolchinsky, P., Blanco, R., Atserias, J., Mika, P., Zaragoza, H. Searching through time in the New York Times. In: HCIR'10 Workshop. pp. 41-44. (2010).
8. Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., Aiden, E.L. Quantitative Analysis of Culture Using Millions of Digitized Books. Science, 331(6014). (2011).
9. Nunes, S., Ribeiro, C., David, G. Use of Temporal Expressions in Web Search. In: ECIR'08. pp. 580-584. Springer-Verlag, (2008).
10. Strötgen, J., Gertz, M. HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions. In: IWSE'10 associated to ACL'10. pp. 321-324. (2010).