Searching through time in the New York Times

HCIR Challenge 2010

Michael Matthews
mikemat@yahoo-inc.com

Pancho Tolchinsky
tolchinsky@gmail.com

Roi Blanco
roi@yahoo-inc.com

Jordi Atserias
jordi@yahoo-inc.com

Peter Mika
pmika@yahoo-inc.com

Hugo Zaragoza
hugoz@yahoo-inc.com

Yahoo! Labs Barcelona
Barcelona, Catalunya (Spain)

ABSTRACT

In this paper we describe the Time Explorer, an application designed for analyzing how news changes over time. We extend on current time-based systems in several important ways. First, Time Explorer is designed to help users discover how entities such as people and locations associated with a query change over time. Second, by searching on time expressions extracted automatically from text, the application allows the user to explore not only how topics evolved in the past, but also how they will continue to evolve in the future. Finally, Time Explorer is designed around an intuitive interface that allows users to interact with time and entities in a powerful way. While aspects of these features can be found in other systems, they are combined in Time Explorer in a way that allows searching through time in no time at all.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Entity Ranking, Information Retrieval

1. INTRODUCTION

The role of time is critical to understanding the news. In current news search engines, time is primarily used to boost the relevance of the most recent stories. While useful when users are interested in the latest news, it may hinder the search experience of those interested in a broader understanding of a particular news story. These users may benefit from a transversal organization of the topic across time so as to better view how the story has evolved and which people and places have shaped the evolution. Furthermore, these users may equally benefit from predictions on how the story might evolve into the future. When searching about a regional conflict, for example, a user should be able to identify what factors lead to the conflict, which people where most influential and when, and how the conflict is likely to evolve in the future. In the following paper, we present Time Explorer[1], a system that has been designed specifically to answer these types of questions. We begin by presenting related work followed by a discussion of the New York Times (NYT) document collection and corpus preparation. We then present the user interface and finally discuss conclusions and ideas for future work.

HCIR 2010 New Brunswick, NJ USA

2. BACKGROUND AND RELATED WORK

Time has long been an integral part of search engine ranking with most major search engine giving a ranking boost for recently published documents, particularly in the news domain. However, recent work [2, 5] has suggested that the time dimension can be further exploited by automatically creating timelines from temporal information extracted from documents both from metadata such as the publication date, but also from temporal expressions found in the text. In recent years, the importance of timelines has been further evidenced as search engines including Google[2] and Cuil[3] have started incorporating timelines into their search results. Work by Baeza-Yates [3] and later by Jatowt et al [6] focus, in particular, on mining collections for statements about future events and provide frameworks for searching into the future. In addition to searching the time dimension, there has been much work in entity search which has the goal returning the entities, such as people and locations, that are most related to a query [4, 1, 7]. As with time search, entity search requires that the entities are either provided as metadata or extracted automatically using named-entity recognition techniques. The primary contribution of our work is to combine these technologies into a working system with an intuitive user interface that allows users to explore the evolution of topics and entities over time in a powerful way.

3. TIME EXPLORER

3.1 Corpus Preparation

The Time Explorer application has been built as part of the European project LivingKnowledge[4] which aims to

[1]Firefox,http://fbmya01.barcelonamedia.org:8080/future/
[2]http://www.google.com/
[3]http://www.cuil.com/
[4]http://livingknowledge-project.eu/

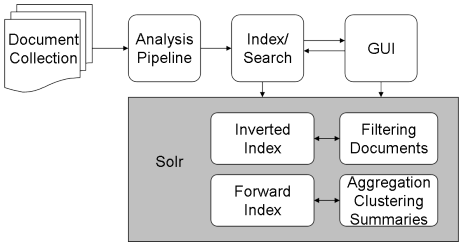


Figure 1: Testbed Architecture

make diversity of knowledge an asset in search applications. The goal is to provide tools that allow exploring knowledge from all points of view and crucially to see how knowledge evolves over time. At the core project is the LivingKnowledge testbed, a planned open source toolkit that allows for annotating document collections with a wide range natural language processing and image analysis tools and provides methods for indexing, searching, and visualizing these annotations using the Solr search engine[5].

The NYT collection of 1.8M news articles from 1987 to 1997 is publicly available, clean, and enriched with high-quality hand-annotated data all of which make it an ideal document collection for evaluating Time Explorer. Though Time Explorer will ultimately aim to incorporate all aspects of diversity covered by LivingKnowledge, the application described here is focused on understanding the time dimension. We have used a subset of the analysis tools available in the testbed including OpenNLP[6] (for tokenization, sentence splitting and part-of-speech tagging, and shallow parsing), the SuperSense tagger[7] (for named entity recognition) and TARSQI Toolkit[8] (for annotating document with TimeML[9]). The resulting analysis is used to extract from each document all of the person, location and organization entities and all time expressions that can be resolved to a specific day, month or year. The time expressions extracted are both explicit as in "September 2010" and relative as in "next month". The relative dates are resolved based on the **publication date** of the article and all dates are associated as **event dates** with the corresponding documents. In addition, simple heuristics are used to assign **keywords** to the document that represent the most important concepts contained in the document, and finally, all of the metadata provided in the NYT collection is associated with each document. From these extractions, two indices are created, one for each document in the collection and one for each sentence in the collection. For the sentence level index, a **content date** is computed as one or more of the **event dates** found in the document or the **publication date** if there are no event dates.

For example, given the following hypothetical document with publication date in May 1^{st}, 1999:

> Slobodan Milošević became president of Yugoslavia in 1997. Slobodan Milošević will run for president again next year.

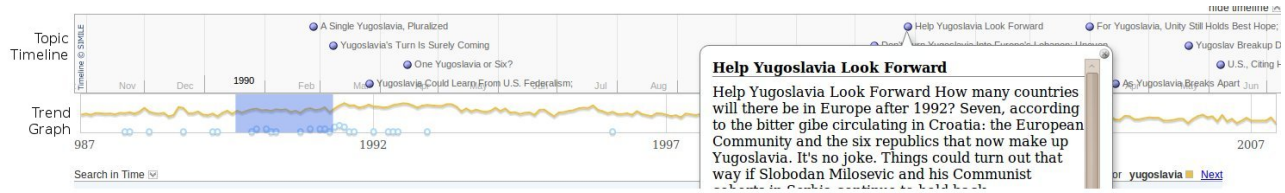
Two sentences will be found. *Slobodan Milošević* will be extracted as a person in both sentences and *Yugoslavia* will be extracted as a location in first sentence. *1997* will be extracted as a time expression in the first sentence and *next year* will be extracted as an expression in the second sentence and resolved to 2000. The **publication date** for both sentences will be May 1^{st}, 1999 while the **content date** of the first sentence will be 1997 and the **content date** of the second sentence will be 2000.

The resulting indices allow for a wide range of queries including: 1) return the documents that contain the word Yugoslavia, 2) return a list of people most related to the query Yugoslavia, 3) return the number of documents containing Yugoslavia that were published in each month, 4) return documents that contain the query Yugoslavia and mention the person Slobodan Milošević, 5) return documents containing the query Yugoslavia that were published in 1999 and 6) return documents containing the query Yugoslavia with events in 2000. These queries and the combinations of them are very powerful but it is unlikely that a user will be able to express the queries in a meaningful way. Therefore defining an intuitive user interface is extremely important.

3.2 User Interface

The focus of this application in understanding how topics evolve over time and thus, not surprisingly, the core of the user interface is a timeline. Though there are many timelines available, including Google trends[10] and Google timeline[11] and many derived from the Simile Timeline widget[12], we attempt to improve on these implementations by combining many of the best features. Figure 2(a) displays the timeline produced for the query "Yugoslavia". The timeline is split between two bands - the bottom band, we call the **trend graph**, shows how the frequency of documents containing Yugoslavia changes over the 20 years covered by the NYT collection while the top band, called the **topic timeline** uses the Simile widget to display the titles of the top ranked articles. As shown, the user can click on the title of the articles to get a document summary. Furthermore, the user is able to scroll through the articles that are displayed in the top window by moving the highlight box with the mouse. Circles indicate which documents are immediately available for viewing. The number of results available for viewing initially is configurable by the application with a trade-off between response time and coverage on the timeline. However, one can easily view more documents for a particular time period by using the mouse to move the highlighted region to a particular point in time and the mouse button to trigger a search. For example, clicking the mouse with the highlight region over the start of the timeline will populate the topic timeline with documents from that time period as shown in Figure 2(b). In this case, this quickly helps one discover that before the conflict started, published articles were dominated by stories of both ethnic and economics problems. When a time period has been selected, the time frame is automatically displayed below the timeline and standard user interface features are used to indicate to the user that they can remove the time range by clicking on the close button, or change the time range manually by entering the dates directly in date fields and selecting search.

[5]http://lucene.apache.org/solr/
[6]http://opennlp.sourceforge.net/
[7]http://sourceforge.net/projects/supersensetag/
[8]http://www.timeml.org/site/tarsqi/
[9]http://www.timeml.org/site/index.html
[10]http://www.google.com/trends
[11]http://www.google.com/
[12]http://www.simile-widgets.org/timeline/



(a) Basic Timeline

(b) Time period Selection

(c) Timeline with Entity trends

Figure 2: Timeline Control

In addition to seeing the articles, an entity list panel displays the entities most associated with the query as shown in Figure 3. The user can view all documents that con-

Figure 3: Entity Selection

tain the entity by clicking on the entity, but can also use a menu to choose to exclude documents containing the entity, submit the entity as a stand alone query and also see a definition of the entity, currently using a simple Wikipedia lookup. These advanced features are only displayed if the user specifically moves the mouse over the icon next to the entity thus keeping the interface simple for the basic user, but still providing useful features in a straightforward manner to the more experienced user.

An additional feature of the entity filter is to provide a trend line for the entity on the trend graph. Figure 2(c) demonstrates this feature. *President Clinton* and *General Wesley Clark* have been selected as entities. The trend lines makes it easy to see when these entities were important with respect to Yugoslavia. *President Clinton* became important when he became president and *General Clark* when he became commander of the NATO forces. Again, there are visual clues as to which entities have been selected and how to remove them from the query if desired. Similar visual clues are also displayed in the entity list as previously shown in Figure 3. The entity list is modified as the query is refined allowing the user to easily see how the important entities change over time for a given query. For Yugoslavia, before the conflict many sports figures were highly associated with query, but, as the conflict progressed, world leaders became much more relevant. Using a similar technique, an evolving relationship was found between Slobodan Milošević and Saddam Hussein. At first, the relationship was largely based on people comparing the relatively unknown Milošević with Hussein and at the end, the relationship was that both were on trial for war crimes. In between, however, a directly relationship was found where Yugoslavia was selling arms to Iraq. In the above scenarios, the timeline is centered around the **publication date** of the articles. However, it is also possible to use the **content date** as the driving date which has the advantage of allowing searching into the future. Figure 4, shows the results for a search on *Iraq*. Using the timeline, it is quick to look into predictions such as the one shown suggesting that Iraq could develop missiles capable of hitting the U.S by 2015. Other predictions include details about the expected cost of the war as well as pre-

Figure 4: Searching the Future

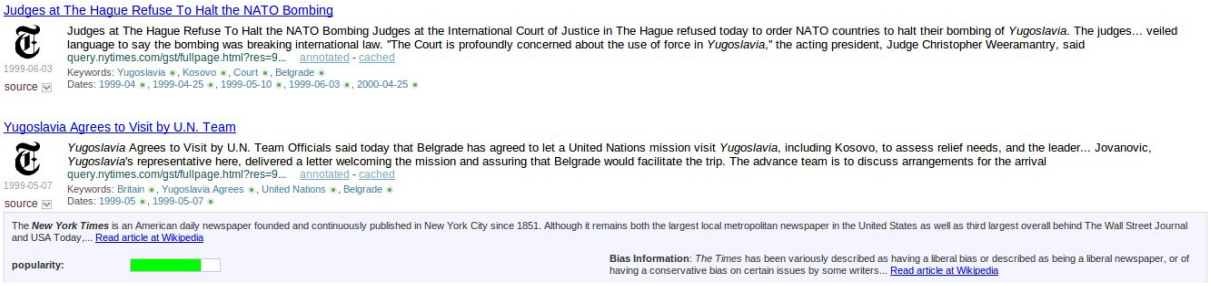


Figure 5: Results

dictions about the success and/or failure on future dates. Using the **content date**, it is also possible to look for articles making predictions about the current date that were made in the past. For example, we were able to look at predictions that were made about 2010 in the articles from the NYT collection. Some results were accurate such as articles discussing a possible 2010 UK election between Gordon Brown and David Cameron, which did take place. Others were amusing like the one from Al Gore during the run-up to the 2000 US presidential election suggesting that his budget proposal would still leave some room for a budget surplus in 2010 - far different from the half-trillion dollar budget deficit actually faced today.

Though the user can learn quite a great deal from the timeline alone, there are also some features in the document snippet shown in Figure 5 that can further assist the user. In addition to the standard highlighted snippet text, there are lists of both the most important keywords associated with the document and the most important dates. These serve to both better summarize the document and to provide an additional mechanism for refining the search. In addition, clicking on the source gives details about the source of the article. In the NYT collection, this is obviously limited to the New York Times, but other collections will include additional news sources and possibly well known authors and bloggers.

4. CONCLUSIONS AND FUTURE WORK

In conclusion, the system presented is an effective tool for analyzing how news topics evolve over time. The application includes many features that, in combination, we believe improve upon what is currently available in news search. Most notably, the tight integration between the trend graph, the topic timeline, and the entity list and the ability to search into the future, but also a user interface which allows for easy query refinement while still providing visual clues that allow the user to understand how he arrived at the current state.

In the future, we plan on integrating LivingKnowledge work on opinion mining and bias detection. The search for the future of Iraq, for example, would be greatly improved if we visualize whether the opinions concerning Iraq are positive or negative and how these opinions change over time and also by visualizing the bias of the opinion holders. We also plan on evaluating the system in a realistic setting to confirm that the system does provide advantages over current technologies.

5. REFERENCES

[1] Overview of the inex 2008 entity ranking track. In *Focused Access to XML Documents: INEX 2008 Dagstuhl Castle, Germany*, 2008.

[2] O. Alonso, M. Gertz, and R. Baeza-Yates. Clustering and exploring search results using timeline constructions. In *CIKM '09*, pages 97–106, New York, NY, USA, 2009. ACM.

[3] R. Baeza-Yates. Searching the future. In *SIGIR Workshop MF/IR*.

[4] K. Balog, A. P. de Vries, P. Serdyukov, P. Thomas, and T. Westerveld. Overview of the trec 2009 entity track. In *TREC 2009 Working Notes*. NIST, November 2009.

[5] R. Catizone, A. Dalli, and Y. Wilks. Evaluating automatically generated timelines from the web. In *LREC 2006*, 2006.

[6] A. Jatowt, K. Kanazawa, S. Oyama, and K. Tanaka. Supporting analysis of future-related information in news archives and the web. In *JCDL '09*, pages 115–124, New York, NY, USA, 2009. ACM.

[7] H. Zaragoza, H. Rode, P. Mika, J. Atserias, M. Ciaramita, and G. Attardi. Ranking very many typed entities on wikipedia. In *CIKM '07*. ACM, 2007.