

Time-Based Exploration of News Archives

Omar Alonso
Microsoft Corp.
Mountain View, CA
omalonso@microsoft.com

Klaus Berberich Srikanta Bedathur Gerhard Weikum
Max-Planck Institute for Informatics
Saarbrücken, Germany
{kberberi, bedathur, weikum}@mpi-inf.mpg.de

ABSTRACT

In this paper, we present NEAT, a prototype system that provides an exploration interface to news archive search. Our prototype visualizes search results making use of two kinds of temporal information, namely, news articles' publication dates but also their contained temporal expressions. The displayed timelines are annotated with major events, harvested using crowdsourcing, to make it easier for users to put the shown search results into context. The prototype has been fully implemented and deployed on the New York Times Annotated Corpus.

Categories and Subject Descriptors

H.3.1 [Content Analysis and Indexing]: Linguistic processing; D.5.2 [Information Interface]: User Interfaces

General Terms

Design, Experimentation, Human Factors

Keywords

Crowdsourcing, timelines, exploration, news archives

1. INTRODUCTION

News archives keep growing in volume and coverage as fresh content is published and old content is being digitized. The New York Times (NYT), as one example, allows users to search and access all of its contents published since 1851. The archive of the British newspaper The Times, as a second example, even goes back until 1785.

When searching news archives, presenting users with a ranked list of few search results is insufficient, as it does not reflect how relevant news articles are spread in the time dimension and thus fails to display the course of history. Instead, it forces users to sift through a large number of relevant news articles and painstakingly piece together how real-world events unfolded.

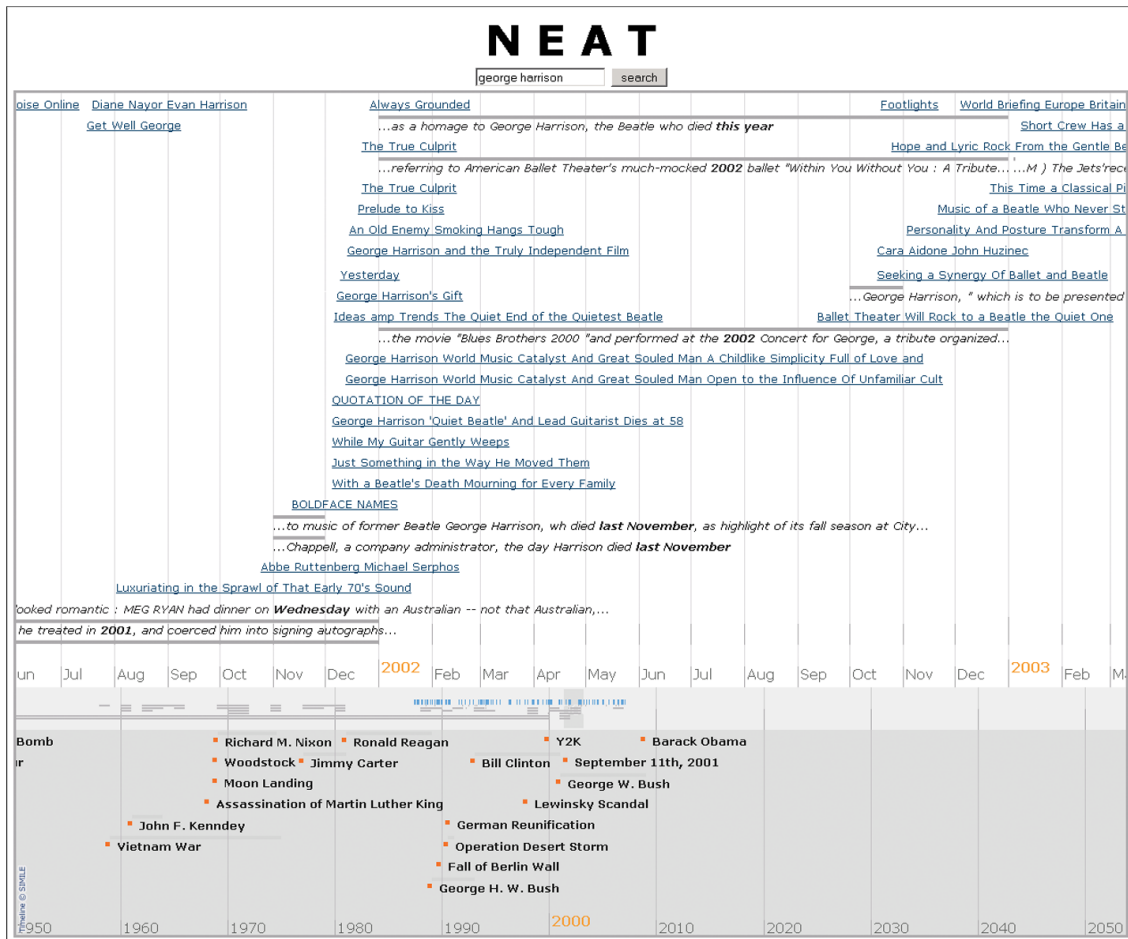
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

HCIR '10 New Brunswick, New Jersey USA
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...\$10.00.

In this paper, we describe our News Exploration Along Time (NEAT) prototype that provides an exploration interface to search news archives. NEAT has been deployed on the New York Times Annotated Corpus [3], as a real-world news archive, and combines several novel features including:

- **Use of Richer Temporal Information:** News articles come with different kinds of temporal information. This includes their publication dates that are typically readily available. However, within the news articles' contents there is often more temporal information hidden. For example, a news article on oil spills published on February 3rd 1991 may contain the following sentences: “*By contrast, the spill caused by the Exxon Valdez in 1989 contained almost 11 million gallons...*” and “*Nearly three years later, he said, young trees are growing in the mangrove...*”. Temporal expressions (e.g., *in 1989*) are another kind of temporal information that can be extracted from the news articles' contents.
- **Snippets with Temporal Information:** NEAT leverages both kinds of temporal information mentioned above. Relevant news articles are anchored on a timeline based on their publication date. Beyond that, NEAT shows relevant snippets that contain temporal expressions and anchors them accordingly. In doing so, NEAT facilitates gaining an understanding of *when relevant news articles were published* but also *which times relevant news content refers to*.
- **Semantic Temporal Anchors:** To aid users in contextualizing the displayed news articles and snippets, our system shows a set of major events that serve as semantic temporal anchors. Examples of such major events include “*Building of Berlin Wall*” (for the year 1961), “*Challenger Disaster*” (for January 1986), and “*Woodstock*” (for the year 1969). We harvest a large collection of such semantic temporal anchors using the crowdsourcing platform Amazon Mechanical Turk. Note that temporal anchors can easily be personalized - users could thus have local libraries of personal anchors (e.g., including their day of birth or wedding day), making it even easier for them to contextualize search results.

Organization. Related work is discussed in Section 2. Section 3 gives details on NEAT's exploration interface. In Section 4, we describe the gathering of timeline annotations using crowdsourcing. NEAT's implementation is subject of Section 5. Finally, in Section 6, we conclude this work and outline next steps.



(a)

(b)

(c)

Figure 1: NEAT screenshot for the query **george harrison** showing (a) main timeline with relevant news articles and relevant temporal snippets, (b) overview timeline, and (c) major events as semantic temporal anchors.

2. RELATED WORK

We now put the present work in context with existing prior research. The “Stuff I’ve Seen” system described by Dumais et al. [9] and similar approaches such as Ringel et al. [13] also make use of temporal information to facilitate information access. However, in their setting, typically only publication dates or timestamps of documents, emails, etc. are considered. In addition we exploit temporal expressions contained in news articles’ contents in our work. The Time Frames system described by Koen and Bender [11] is similar to our work, since it also uses temporal expressions contained in news articles. Their main focus, though, is on supporting users in reading news articles, but not on search and exploration.

Our own earlier work is also related but focuses on different aspects. Alonso et al. [7] present an approach for clustering and exploring search results in timelines. Berberich et al. [8] describe a model for temporal information needs that makes use of temporal expressions. Both approaches use crowdsourcing for their respective evaluations.

Other related research includes the recently proposed Meme-tracker system [12] that tracks the mutational flow of so-called memes over time. Their system, though, focuses on pre-identified memes and does not support arbitrary ad-hoc

queries. Jones and Diaz [10] show that the temporal profile of a query, determined based on the publication dates of relevant documents, is useful in query classification. Swan and Allan [15], as an early piece of research, focus on automatically generating overview timelines for a collection of documents. Wang and McCallum [17] is a more recent, more sophisticated approach along similar lines. It is conceivable to augment NEAT with such topical overviews.

Google has recently added the `view:timeline` feature to display search results along a timeline. Similarly, Google News Archive Search [2] also visualizes the query results as a temporal frequency distribution of relevant documents. While such visualization provides a high-level view of the topical popularity, they do not makes use of temporal expressions contained in documents and thus do not provide interesting snippets corresponding to a time period. Finally, TimeSearch [5], another related prototype, also makes use of temporal expressions contained in relevant documents.

3. TEMPORAL EXPLORATION

We now describe NEAT’s exploration interface in more detail. Figure 1 shows a screenshot of the interface when displaying results for the query **george harrison**. In detail, the interface consists of the following timelines:

NEAT

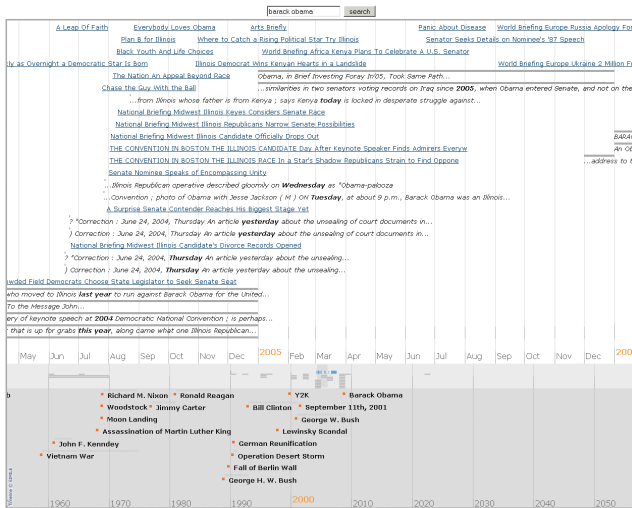


Figure 2: Results for barack obama around 2005

- (a) **Main timeline** showing titles of relevant news articles (e.g., “The True Culprit”) placed according to their publication date and relevant temporal snippets (e.g., “...as a homage to George Harrison, the Beatle who died this year”) placed according to their contained temporal expressions.
- (b) **Overview timeline** summarizing relevant news articles and temporal snippets shown in (a) at a coarser temporal granularity.
- (c) **Major events**, gathered using crowdsourcing as described in more detail in Section 4, that serve as semantic temporal anchors for the users.

Notice that the timelines are synchronized, so that navigating in one will automatically adjust the others.

We distinguish two time dimensions in NEAT, namely, publication time and reference time. By placing titles of relevant news articles on the timelines based on their publication time (i.e., when they were published), we provide users with an overview of relevant news articles and the order of real-world events behind them. Reference time, as the second time dimension considered, reflects which times relevant news content refers to. To illustrate the difference between publication time and reference time, consider an article published in June 2010 that compares this year’s FIFA World Cup against earlier instances of the tournament. Whereas the article’s title would be placed on the day of its publication in June 2010 according to publication time, parts of its content, so-called temporal snippets, would be placed, for instance, at the years 2006, 2002, 1998 etc. depending on which earlier FIFA World Cup they talk about. As the example suggests, in order to get a hold on the reference-time dimension, we must identify the times that an article’s content refers to. This can be accomplished using existing tools for identifying and interpreting temporal expressions, such as TARSQI [16] or TimexTag [6], that are readily available. By showing relevant temporal snippets, we provide the user with a means to explore the content of many documents at once, which is less time-consuming than sifting through each of them separately.

NEAT

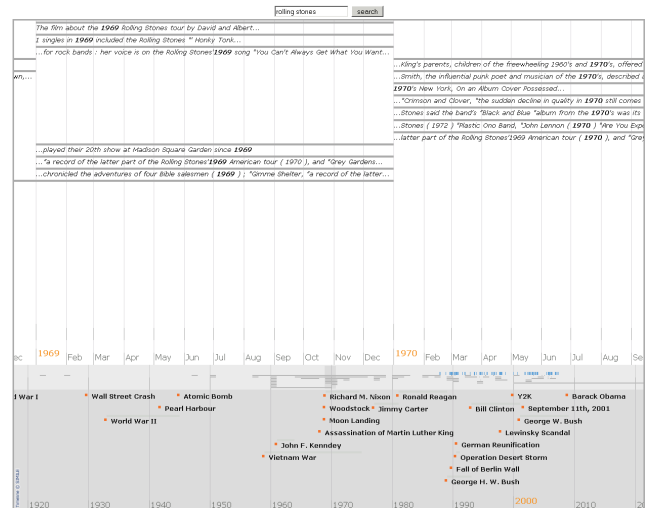


Figure 3: Results for rolling stones around 1970

Figure 2 and 3 show two more anecdotal examples of NEAT in action. As shown in Figure 2, for the query **barack obama**, it is apparent from the overview timeline that there is little relevant content before 2005 – the year when Barack Obama became United States Senator. For the query **rolling stones**, as our second example shown in Figure 3, we see that, by showing relevant temporal snippets, NEAT offers insights into the rock band’s activities during the 1970s, which is long before the publication dates of news articles in the New York Times Annotated Corpus. Apart from that, for both examples, the major events shown provide an interesting political and societal context.

4. TIMELINE ANNOTATION USING CROWDSOURCING

An important item that arises when working with timelines is the selection of the main events and how they should be presented. In the particular case of a newspaper like NYT that contains a wealth of information, how do we select the most representative events? An obvious approach would be to select the events based on coverage or popularity. However, the quality of the timeline in this case would be purely based on the newspaper’s content. Instead, we took a “wisdom of the crowd” approach. The idea is to annotate the timeline based on collective human knowledge. We model this as a bipartite graph where we want to match a temporal expression to an event. We believe this may provide a more realistic representation of major events. We gathered temporal annotations at large-scale using Amazon Mechanical Turk (AMT) [1]. In a series of experiments, each HIT (Human Intelligence Task) on AMT consists of a request to expand a temporal expression with an entity (e.g., a person, country, or organization) or event. Based on the agreement level among workers, we derive key entities for constructing a semantic temporal annotation layer on top the timeline. The outcome is a manually annotated timeline that helps users in contextualizing anchor search results. We paid \$0.01 per assignment and each task was completed by five different workers. We manually created a set of 50 temporal expressions that represent time at different granularities as follows:

- Dates (e.g., *9/1/1939* or *4/4/1968*)
- Relative (e.g., *last year*, *next year*, or *tomorrow*)
- Weekdays (e.g., *Monday* or *Tuesday*)
- Months (e.g., *January* or *February*)
- Years (e.g., *1492*, *1945*, or *1970*)
- Decades (e.g., *60s*, *70s*, or *80s*)
- Centuries (e.g., *19th* or *20th*)

We ran the experiment for different categories (politics, sports, culture, world affairs, movies, and music) using the same set of temporal expressions. By analyzing the data we can see that an explicit temporal expression tends to have a clear annotation as we see in the following examples verbatim. In the case of “1492”, the workers wrote: America, Christopher Columbus, Columbus, Columbus discovers America, France. For relative expressions, the annotation tends to be of less value. For the temporal expression “4pm”, we have: Afternoon, Bakers, Mauritius, Oprah Winfrey, TED. Going at a higher level than year, decades also provide interesting information. For example, for “70s”, we have: disco, oil shocks, Richard Nixon, usa, Watergate. Months provide a mix of typical calendar events with some other observations. For “March”, workers wrote: Brutus killed Julius Caesar on the ides, caesar, Easter, saint patrick, St. Patrick’s Day. The next step is to get a consensus among workers and select one or two significant events for that particular temporal expression. Examples of annotations produced by crowdsourcing are (1969: Woodstock, Moon landing), (1970: Nixon), and (2003-2009: Iraq war) to name a few with different time granularities. It is not always possible to get consensus on an <event, temporal expression> pair. An interesting example is the year “1982”, where the crowd annotated: Ronald Reagan, Spain World Cup, Charles & Lady Di wedding, and Falklands War. These are all valid events and probably interesting on their own, but we were not able to find consensus on one or two.

5. IMPLEMENTATION

We now provide some details on the implementation of our NEAT prototype. Prior to indexing the dataset using our prototype, we annotated temporal expressions using TARSQI [16]. To implement the user interface, shown in Figure 1, we make use of the timeline visualization provided as part of the SIMILE project [4]. When the user issues a query, a request is sent to a Java servlet. This Java servlet, running in the backend, then processes the user query by retrieving a fixed number of relevant documents and a fixed number of relevant temporal snippets. Notice that the retrieved temporal snippets are independent from the retrieved relevant documents, thus fostering diversity of displayed information. To retrieve relevant documents and temporal snippets, the servlet accesses two inverted indexes, one for documents and one for snippets, that are implemented using an Oracle 11g database. To determine the relevance of news articles and temporal snippets, we employ Okapi BM25 [14] as a retrieval model. For temporal snippets, we slightly modify the retrieval model, using the number of temporal expressions contained in a snippet as a multiplicative boosting factor. Finally, before sending a response, the servlet looks up metadata for the identified

documents and phrases (e.g., their URLs and publication dates), and adds markup to highlight query terms and temporal expressions.

6. CONCLUSIONS AND FUTURE WORK

We presented NEAT, a working prototype for exploring news along timelines. We used the New York Times Annotated Corpus to show the features of our system. The prototype is easy to use and the authors found it interesting to navigate to the past when issuing queries about current world affairs.

Future work includes a user study of the user interface to get a better idea (and metrics) about the prototype. Previous research [7] has shown that users like to see information in time, so we would to explore this in more detail. The annotation of timelines by major events gathered using crowdsourcing looks very promising. A limitation is that the annotation depends a lot on the quality of the workers and, in our experience, the annotations seemed to have an American flavor instead of being world representative. We plan to keep working on this aspect.

7. REFERENCES

- [1] Amazon Mechanical Turk <http://www.mturk.com>.
- [2] Google News Archive Search <http://news.google.com/archivesearch>.
- [3] New York Times Annotated Corpus <http://corpus.nytimes.com>.
- [4] SIMILE TimeLine Visualization. <http://simile.mit.edu/timeline/>.
- [5] TimeSearch History <http://www.timesearch.info>.
- [6] D. Ahn et al. A Cascaded Machine Learning Approach to Interpreting Temporal Expressions. *HLT-NAACL 2007*.
- [7] O. Alonso et al. Clustering and exploring search results using timeline constructions. *CIKM 2009*.
- [8] K. Berberich et al. A Language Modeling Approach for Temporal Information Needs. *ECIR 2010*.
- [9] S. T. Dumais et al. Stuff I’ve seen: a system for personal information retrieval and re-use. *SIGIR 2003*.
- [10] R. Jones and F. Diaz. Temporal profiles of queries. *ACM TOIS 2007*.
- [11] D. B. Koen and W. Bender. Time frames: Temporal augmentation of the news. *IBM Systems Journal 2000*.
- [12] J. Leskovec et al. Meme-tracking and the dynamics of the news cycle. *KDD 2009*.
- [13] M. Ringel et al. Milestones in Time: The Value of Landmarks in Retrieving Information from Personal Stores. *INTERACT 2003*.
- [14] S. E. Robertson and S. Walker. Okapi/keenbow at trec-8. 1999.
- [15] R. Swan and J. Allan. Automatic generation of overview timelines. *SIGIR 2000*.
- [16] M. Verhagen et al. A. Rumshisky, J. Phillips, and J. Pustejovsky. Automating Temporal Annotation with TARSQI. *ACL 2005*.
- [17] X. Wang and A. McCallum. Topics over time: a non-Markov continuous-time model of topical trends. *KDD 2006*.