Exploratory Search Using Timelines

Omar Alonso
Department of Computer Science
University of California, Davis, CA
oralonso@ucdavis.edu

Ricardo Baeza-Yates
Yahoo! Research
Barcelona, Spain
rbaeza@dcc.uchile.cl

Michael Gertz
Department of Computer Science
University of California, Davis, CA
gertz@cs.ucdavis.edu

ABSTRACT

As search applications keep gathering new and diverse information sources, presenting relevant information anchored in time becomes more important. Temporal information is available in every document either explicitly, e.g., in the form of temporal expressions, or implicitly in the form of metadata. Recognizing such temporal information and exploiting it for document retrieval and presentation purposes are important features that can significantly improve the functionality of search applications. In this paper, we present an exploratory search interface that uses timelines to present and explore search results. We also describe a prototypical implementation that illustrates the main ideas of our approach.

Author Keywords

Temporal expressions, temporal document retrieval, user interface, visualization.

ACM Classification Keywords

H.3.3 [Information Search and Retrieval]: Clustering, Retrieval Methods. H.5 [Information Interfaces and Presentation]: General

INTRODUCTION

Current interfaces to search engines typically present search results ordered by the relevance of documents from a document collection to a search query. For this, the freshness of the information, that is., documents or parts thereof, is considered an important part of the result quality. Temporal attributes in Web pages or documents such as date, however, are just viewed as some structured criteria to sort the results in descending order of relevance.

In search situations where the task requires the browsing and exploration of a search result [11], we argue that temporal information can help significantly to accomplish respective tasks. The presentation of relevant information along a well-defined and understood timeline is an important step to find, for example, the most recent document relevant to a query or the first point in time a document (based on the temporal information contained in the document) is relevant to the query.

Our approach supports the exploitation of temporal information in documents, and the usage of such information to anchor search results along a well-defined timeline. We believe that such timelines should be an essential part of every exploratory document search system.

Research in using time for retrieval and browsing activities is fairly recent. The use of tags to visualize photos taken over a period of time is a good example of how useful time can be for arranging objects [4]. Furthermore, in addition to topic detection and tracking, the discovery of bursts in a stream of content can be useful for the identification of topics [6]. In particular, in settings where a user is looking for relevant documents in a less familiar domain, we would like to show peaks of activity (in the form of documents) over time that contain information relevant to the user query.

We agree with previous work that placing search results in a timeline can facilitate the exploration of information [1], [7], [8]. Our approach, however, differs in a number of ways. We do not restrict the search space to a personal desktop environment, because we believe that timelines should be an integral part of search applications. Ringel et al. [8] concentrate on “object” timestamp (e.g., a document, email message, or presentation), whereas we use both temporal expressions and document metadata as document timestamps alone can often be misleading. Exploiting the temporal expressions embedded within a document leads to a much richer framework for search result exploration.

TEMPORAL EXPRESSIONS

A document typically has temporal metadata, such as the creation date or modification date. We also observe that the content of a document typically has *temporal references* to past and/or future points in time. These temporal references are either (1) explicitly represented, such as a date and time in a calendar (“March 12, 2004”), (2) denoted as events that have an associated time value, such as a holiday

("Christmas" or "Thanksgivings"), or (3) represented by a vague time reference ("by Friday").

The detection and extraction of time information from documents utilizes *named-entity extraction* techniques [5]. The output of this document preprocessing step is a time-annotated document, where the desired time information and expressions are represented in a specific format, for example, in the form of XML, possibly outside the original document [10]. Named-entity extraction is a very scalable approach, e.g., for the identification of holidays, and does not just rely on document specific time mappings as employed, e.g., in an email-calendar based approach [7].

We illustrate the above ideas of exploiting temporal information associated with documents using a literature search scenario based on the DBLP bibliography data set [3], which contains detailed information about journal, conference, and workshop publications, in particular the date of each publication.

TIMELINE CONSTRUCTION

Time has been a subject of study in many disciplines, and it is usually represented as a continuum line with origin and no end. We assume a time representation based on the Gregorian calendar, with a single day being an atomic time interval called *chronon*. A timeline consists of a sequence of chronons, and optionally can have a start and/or end chronon. Consecutives chronons can be grouped into more coarse-grained time *granules* such as weeks, months, years, etc.

As stated before, our goal is to present search results that are arranged in a timeline. There are several ways to accomplish this. The most obvious approach is to use the document metadata for anchoring documents in a timeline. Unfortunately, if the time range is large, i.e., the documents' metadata are widespread over a long time interval, the timeline might be too fine-grained and therefore too large. It thus makes more sense to group a search result based on a more general, coarse-grained time granule, such as year, and if needed, allow the user to explore a particular year chronons at a more fine-grained level, e.g., at months, weeks, or days.

The timeline construction is based on a clustering algorithm that uses temporal expressions extracted from documents and anchoring these expressions in a timeline [2]. The first step is to match the query terms with the document text near temporal expressions. Since our document collection consist of research articles, all of them having publication year as a temporal expression anyway, the relevancy match is reduced to whether or not the query terms occur in the document.

The search engine retrieves the results by relevancy using a traditional tf/idf metric. The resulting hit list is then clustered by year and within each year, documents are ranked by score.

The timeline is made of clusters (labeled by year), and also provides for more fine-grained cluster exploration for each year. It is important to note that if a document has more than one temporal expression, it can appear in more than one cluster, and it can also be ranked differently in different clusters.

PROTOTYPE IMPLEMENTATION

The document collection consists of approximately 297,000 journal papers records from the DBLP bibliography data set. Since the data is available in XML format, we loaded the data into an Oracle database for storage. The extensions to SQL allow us to use XPath and search features for querying the text indexes.

For the visualization part, we use the SIMILE timeline AJAX toolkit [9], for presenting time-based events. The exploratory search prototype is a Web-based application and no specific plug-in is required.

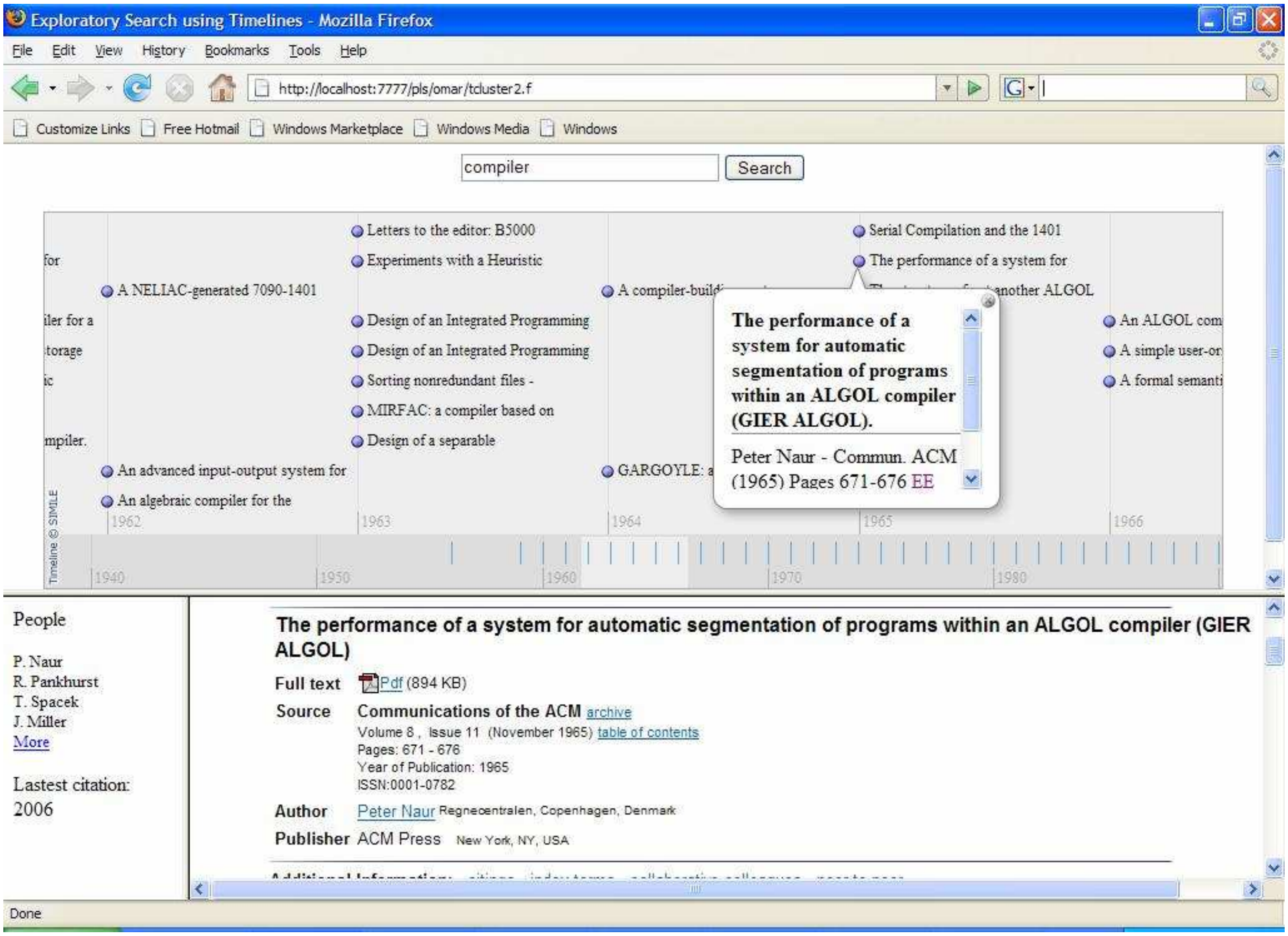
The interface is organized as follows. The main section takes half of the screen and contains the search box and the timeline. The timeline consists of two *bands* that represent different time scales: decade and year. Both bands are synchronized such that panning one band also scrolls the other. The lower band (decade) is much smaller since the goal is to show activity in a decade. The upper band shows all articles in a given year. When the user clicks on an item bullet, the bibliographical information is presented (e.g., author, title, journal, etc.). If the user clicks on the "EE" link (electronic copy), the article content is presented in a separate frame.

Figure 1 shows the exploratory search interface in action for the query results about "compiler". The system retrieves all journal articles that contain "compiler" in the title and returns a hit list clustered by year. All the search results are anchored in the timeline. If more than one article falls within a year, the order is based on its relevance to the query. In this example, we can use the lower band to pick a decade. We can move to the 50s where the early papers on compilers where published or stay in the current year to see the latest publications. Say we are interested in papers from the 60s, because they were very influential. After selecting that particular decade, we can see an interesting number of articles including the one selected about GIER ALGOL.

In summary, the user can see all search results in a timeline, observe years of high activity (i.e., many document are relevant to the query terms in respective years), explore the items using both bands and then click to select a particular article. A separate frame shows the complete article as well as other information such as related people and the latest year of a citation.

This output of the search results is an intermediate representation, which is transformed to the timeline visualization format. This is useful in case one wants to use other timeline visualizations. The toolkit is flexible enough so it is also possible to show the same information in a vertical timeline.

Compared to other literature search systems (DBLP, Google Scholar, etc.), our prototype system allows the user to see the presence of articles in time as well as years of high activity. We believe this type of representation of high activity is crucial for domains where there is a concentration on the research or analysis part.



for

iler for a

torage

ic

mpiler.

A NELIAC-generated 7090-1401

An advanced input-output system for

An algebraic compiler for the

Letters to the editor: B5000

Experiments with a Heuristic

Design of an Integrated Programming

Design of an Integrated Programming

Sorting nonredundant files -

MIRFAC: a compiler based on

Design of a separable

A compiler-build

GARGOYLE: a

Serial Compilation and the 1401

The performance of a system for

another ALGOL

An ALGOL com

A simple user-or

A formal semanti

The performance of a system for automatic segmentation of programs within an ALGOL compiler (GIER ALGOL).

Peter Naur - Commun. ACM (1965) Pages 671-676 EE

People

P. Naur
R. Pankhurst
T. Spacek
J. Miller
More

Lastest citation:
2006

The performance of a system for automatic segmentation of programs within an ALGOL compiler (GIER ALGOL)

Full text Pdf (894 KB)

Source Communications of the ACM archive
Volume 8 , Issue 11 (November 1965) table of contents
Pages: 671 - 676
Year of Publication: 1965
ISSN:0001-0782

Author Peter Naur Regnecentralen, Copenhagen, Denmark

Publisher ACM Press New York, NY, USA

Figure 1. Exploring articles about compilers in a timeline, which is formed by two bands that represent decade and year respectively.

CONCLUSIONS

Adding time as part of exploratory search tasks can lead to interesting discoveries. We propose to use well-defined timelines as an alternative view for presenting search results that have rich temporal expressions or where the usage of time plays an important role.

Future work includes adding more implicit time information to the timeline visualization and to enhance its presentation with more metadata.

For a user study and evaluation, our initial plan is to assign a number of search tasks to participants and compare the results using our approach and existing bibliographic systems on the Web. For example, finding the first article on a topic or the period of time where most of the research for a particular sub-area has been done. We would also like to see the influence of time in search results rather than just popularity or last modified date attributes.

REFERENCES

1. Allen, R. "A Focus-Context Browser for Multiple Timelines", *JCDL* 2005, Denver, CO.

2. Alonso, O. and Gertz, M. “Clustering of Search Results using Temporal Attributes”, *SIGIR* 2006, Seattle, WA.
3. DBLP Computer Science Bibliography. http://dblp.uni-trier.de
4. Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P., and Tomkins, A. “Visualizing Tags over Time”. *WWW* 2006, Edinburgh, UK.
5. Jackson, P. and Moulinier, I. *Natural Language Processing for Online Applications*. John Benjamins (2002).
6. Kleinberg, J. “Bursty and Hierarchical Structure in Streams” *8th ACM SIGKDD*, 2002.
7. Krishnan, A. and Jones, S. “TimeSpace: activity-based temporal visualization of personal information spaces”, *Personal and Ubiquitous Computing* 9(1): 46-65 (2005).
8. Ringel, M., Cutrell, E., Dumais, S, and Horvitz, E. “Milestones in Time: The Value of Landmarks in Retrieving Information from Personal Stores”, *Proceedings of Interact 2003,* Zürich, Switzerland.
9. TimeLine http://simile.mit.edu/timeline/
10. TimeML http://timeml.org
11. White, R., Kules, K., Drucker, S., and schraefel m. (Eds). “Supporting Exploratory Search”, *CACM*, Vol. 49, No. 4, April 2006.