PAPER

On the Value of Temporal Information in Information Retrieval

Omar Alonso[1], Michael Gertz[1] and Ricardo Baeza-Yates[2]
[1]Department of Computer Science
University of California at Davis, Davis, CA.
[2]Yahoo! Research Barcelona, Spain
oralonso@udavis.edu, gertz@ucdavis.edu,
rbaeza@dcc.uchile.cl

Abstract

Time is an important dimension of any information space and can be very useful in information retrieval. Current information retrieval systems and applications do not take advantage of all the time information available in the content of documents to provide better search results and user experience. In this paper we show some of the areas that can benefit from exploiting such temporal information.

1 Introduction

As the amount of generated information increases so rapidly in the digital world, the concept of time as dimension along which information can be organized and explored becomes more and more important. Time and time measurements can help in recreating a particular historical period or describing the context of a document or document collection, which then can be helpful for relevancy ranking purposes.

As an alternative to document ranking techniques like those based on popularity, time can be valuable for placing search results in a timeline for document exploration purposes. Current information retrieval systems and applications, however, do not take advantage of all the time information available within documents to provide better search results and thus to improve the user experience.

A quick look at any of the current search engines and information retrieval systems shows that the temporal viewpoint is restricted to sorting the search result represented in a hit list by date only. The date attribute is mainly the creation or last modified date of a Web page or document. In some cases, this can be misleading, because the timestamp provided by a Web server or any other document management system may not be accurate. Other search applications provide a range date search as part of the advanced search options. Still, the search results are filtered based on the date attribute. For search purposes, the time axis is mainly constructed using that type of document metadata.

Even simple queries against Web search engines show that oftentimes organizing the documents in a hit list along some timeline can be helpful. For example, a query for “soccer world cup” against search engines now returns mostly pointers to documents that cover the recent event in Germany. But every soccer fan knows that this event happens every four years. Clearly, it would be useful if a tool on top of a traditional retrieval system is more aware of the temporal information embedded in the documents and allows the user to have search results presented in different ways based on the temporal information. For this, it is essential to extract temporal information from documents and associate documents with points in time along well-defined timelines.

1.1 Motivating Examples

One can argue that for certain types of data sources or depending on a particular application, the role of time is implicit in the query. If one is looking for information about “Q3 earnings” for a company, it is assumed that the query means information that pertains to the current fiscal year.

This is typical for content that has a short time span such as news articles. News is essentially new information or current events. If the news is old, it is generally assumed that it is not relevant. On the other hand if the query is “tsunami disaster”, and given that the event happened a few years ago, one expects to retrieve the fact of the event, not necessarily the latest news (although that can be the case, too). Another example is a query about the Iraq war conflict, where the results are based on the latest events with little content from the early 1990s war. In short, time does impact the quality of search results. Ideally, we would like a search engine to be aware of the temporal information embedded in documents and present the results in a time context.

In contrast to the Web where crawlers continuously run to collect new content and eliminate references to broken links, Intranet search engines are starting to go into the opposite direction. For example, based on the Sarbanes-Oxley legislation where public companies are required to keep records of assets (email, for example) for periods of time, it is becoming important to store every reference to a document - even if the document is no longer present. In case of an audit, forensic retrieval (searching the past) can be performed and derive the existence of such a document from the index. A query like “sell stock before earnings” in a particular timeframe determines its relevancy. The important scenario here is “what did you know at what period of time?”

There are other vertical search applications like those in Health Care, where certain data sets such as patient discharge summaries contain time information. In reconstructing a patient’s medical history, the ability to find events and present them in a timeline is a key aspect for establishing the accuracy of the report.

From a user’s perspective, the relevancy of a query has a temporal aspect. The more data sources an information retrieval system acquires, the more important the temporal aspect can be in the retrieval process. Instead of assuming that the user wants relevant search results implicitly sorted by date, it would be interesting to investigate a system that is aware of time for relevancy and shows search results in a temporal context. It can also be useful to filter the “trendy stuff” from the rest.

Every retrieval engine combines search and browsing features by categories (predefined or generated). One can also extend the notion of browsing by adding temporal attributes. A clear advantage is that there is a predefined order for temporal items. For example, Wednesday is before Thursday and the year 2002 describes the range of days from January 1st to December 31st. All of us are used to using a calendar metaphor to manage and view main events in time. A combination of calendar and browsing looks like an interesting alternative to searching and navigating large collections from a temporal perspective.

2 What is Time?

Time has been the subject of study in many disciplines, particularly in philosophy, physics, logic, and Art. We start the study of time information in documents with a number of definitions to set up the right context. A look at an English dictionary shows the following definition for "time": a) a non spatial continuum in which events occur in apparently irreversible succession from the past through the present to the future; b) An interval separating two points on this continuum; a duration.

Since the notion of *event* appears quite often in combination with the notion of time, its dictionary definition says: Event: a) Something that takes place; an occurrence; b) A significant occurrence or happening; c) A social gathering or activity.

Time is defined operationally and involves the process of measuring the units chosen such as millisecond, minute, day, and century. A *timeline*, also known as a chronology, is a linear representation of events in the order in which they occurred. This sequence of events is usually arranged in chronological order and presented in a line display drawn left to right or top to bottom.

A *calendar* is a human designed system for representing physical time. A calendar defines the time values, called *granularities*, of interest to a user, usually over a specific segment of the timeline. One calendar familiar to many is the Gregorian calendar.

2.1 Time and Timelines

As the basis for associating points in time with documents, it is customary to assume a discrete representation of time based on the Gregorian calendar, with a single day being an atomic time interval called *chronon*. A base timeline, denoted T_d, is an interval of consecutive day chronons. For example, the sequence "March 12, 2002; March 13, 2002; March 14, 2002" is a contiguous subsequence of chronons in T_d. Contiguous sequences of chronons can be grouped into larger units called granules, such as weeks, months, years, or decades. A grouping based on a granule results in a more coarse-grained timeline, such as T_w based on weeks, T_m based on months, or T_y based on years. Examples of week chronons in T_w are "3rd week of 2005" or "last week of 2006". Depending on the type of underlying calendar, base timeline, and grouping of chronons, timelines of different time granularity can be constructed.

2.2 Temporal Expressions

There is quite a lot of temporal information in any corpus of documents. Besides a simple document timestamp, what types of temporal information are there and how do they relate to timelines? A *temporal entity* describes a point in time, event, or time period at a conceptual level. The identification of such entities involves a linguistic analysis of the document, where approaches based on named-entity extraction determine *temporal expressions*. A temporal expression is basically a sequence of tokens that represent an instance of a temporal entity.

Similar to the approach by Schilder and Habel [10], here we identify the following three categories:

- Explicit. These temporal expressions directly describe entries in some timeline, such as an exact date or year. For example, the token sequences "December 2004" or "September 12, 2005" in a document are explicit temporal expressions and can be mapped directly to chronons in a timeline.
- Implicit. Depending on the underlying time ontology and capabilities of the named entity extraction approach, even apparently imprecise temporal information, such as names of holidays or events can be anchored in a timeline. For example, the token sequence "Columbus Day 2006" in the text of a document can be mapped to the expression "October 12, 2006", or the sequence "Labor Day 2008" can be mapped to "September 1, 2008".
- Relative. These temporal expressions represent temporal entities that can only be anchored in a timeline in reference to another explicit or implicit, already anchored temporal expression (which, in the worst case, is the document timestamp). For example, the expression "today" alone cannot be anchored in any timeline. However, it can be anchored if the document is known to have a creation date. This date then can be used as a reference for that expression, which then can be mapped to a chronon. There are many instances of relative temporal expressions, such as the names of weekdays (e.g., "on Thursday") or months (e.g., "in July") or references to such points in time like "next week" or "last Friday".

3 Extracting Time from Documents

The extraction of temporal expressions from documents can be accomplished using an approach similar to named-entity extraction. We call such an approach a document annotation pipeline. The first step is to extract time metadata from the document. This can be the creation or last modified date of a file. In case of a Web page, we rely on the information provided by the Web server. The second step is to run a part of speech tagger (POS tagger) on every document. A POS tagger returns the document with parts of speech assigned to each word/token like noun, verb, etc. The tagger also tags sentences delimiters that later are needed for temporal annotation. The third step is to run a temporal expression tagger like GUTime on the POS-tagged version of the document [5]. This step extracts temporal expressions based on the TimeML standard and produces an XML document. We rely on the TimeML specification for temporal annotation, because it has emerged as the standard markup language for events and temporal expressions in natural language [13]. Other approaches can be found in [11].

4 Applications

In this section, we outline a few application areas that can benefit from using temporal information in different ways. The list is by no means complete and the purpose is to only highlight how time can provide new insights. Earliest work on using temporal information are [1,6]. Another look at time from the linguist and natural language aspect can be found in [7,8].

4.1 Ad-hoc retrieval

An information retrieval system helps users find documents that satisfy their needs. We all know that users are not very expressive in what they want and that the information they provide can be ambiguous. The central idea in *temporal information retrieval* is to utilize the temporal expressions that have been determined for each document in a given document collection in order to rank search results based on the temporal information embedded in the documents. By using this approach, time plays a central role in the overall quality of search results. The search application retrieves the result based on the relevance of the documents with respect to the query using traditional metrics *and* the distance of the query terms to temporal expressions in the documents. After all, tense happens at the sentence level so it is important to detect these "boundaries" with respect to the query.

4.2 Hit-list clustering

Hit-list clustering has emerged as an alternative mechanism to present similar documents without requiring the user to go through hundreds of items. Clustering of result documents can lead to better user interfaces and, therefore, to an improved user experience, in particular in the context of information exploration. There are several ways in which the temporal expressions in the set of retrieved documents can be used to group the documents using temporal aspects. Given such a hit list, the first step is to construct a time outline for the documents in the hit list. The documents are then clustered along this timeline, again based on their temporal characteristics. A single cluster can be thought of as a bin that contains only documents with temporal expressions matching the cluster label, which corresponds to some chronon from a timeline. The organization of clusters along a timeline as well as the lattice structure imposed among timelines then allows for the exploration of document clusters at different levels of time granularity [2].

4.3 Exploratory Search

Exploratory search systems have emerged as a specialization of information exploration to support serendipity, learning, and investigation of large data sets. In search situations where the task requires browsing and exploration of a search result, we argue that temporal information can help significantly to accomplish respective tasks. The presentation of relevant information along a 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 [3,9]. The use of time and timelines for clustering and browsing nicely fits exploratory search systems that go beyond simply returning some documents or answer in response to a query.

4.4 Presentation

Current interfaces to search engines typically present search results sorted by the relevance of documents from a document collection to a search query. Temporal attributes in Web pages or documents such as date, however, are just viewed as some structured criteria to sort the result in descending order of relevance. Traditionally, time is represented by an arrow where events are place along the timeline. Recently, new visualization tools have emerged with alternative variations. For example, TimeWall allows a user to define *cards* with attributes over long horizons while being able to focus in on a particular time of interest [14]. SIMILE allows the construction of a timeline that can contain one or more time *bands*, which can be panned infinitely by dragging with the mouse. A band can be configured to synchronize with another band such that panning one band also scrolls the other [12]. Finally, Google has added the `view:timeline` mechanism that allows to see certain search results presented in a timeline [4].

5 Conclusions

Temporal information embedded in documents in the form of temporal expressions provides an import means to further enhance the functionality of current information retrieval applications. Recognizing such temporal information and exploiting it for document retrieval and related applications are important features that can significantly improve the current functionality of search applications. We believe that when a user is engaged in tasks that require time-related investigation and sensemaking, traditional information retrieval and search engines fall short if they do not fully exploit the various types of temporal information embedded in documents.

6 References

[1] James Allan, Rahul Gupta, and Vikas Khandelwal. “Temporal Summaries of News Topics”. In *Proc. of the 24th SIGIR*, 10–18, 2001.

[2] Omar Alonso and Michael Gertz. Clustering of search results using temporal attributes. In Proc. of the 29th SIGIR, 597–598, 2006.

[3] Omar Alonso, Ricardo Baeza-Yates, and Michael Gertz. Exploratory Search Using Timelines. In SIGCHI 2007 Workshop on Exploratory Search and HCI Workshop, 2007.

[4] Google Timeline, http://www.google.com/experimental/

[5] GUTime, http://complingone.geogretown.edu/~linguist

[6] Douglas B. Koen and Walter Bender. “Time Frames: Temporal augmentation of the news”. *IBM Systems Journal* 39(3&4), 2000.

[7] Inderjeet Mani, James Pustejovsky, and Beth Sundheim. “Introduction to the Special Issue on Temporal Information Processing”. *ACM Transactions on Asian Language Information Processing*, 3(1):1–10, March 2004.

[8] Inderjeet Mani, James Pustejovsky, J., and Robert J. Gaizauskas R. (Eds.). *The Language of Time*. Oxford University Press, 2005.

[9] Meredith Ringel, Edward Cutrell, Susan T. Dumais, Eric Horvitz. “Milestones in Time: The Value of Landmarks in Retrieving Information from Personal Stores”. *IFIP TC13 Intern. Conf. on Human-Computer Interaction*, 184–191, 2003.

[10] Frank Schilder and Christopher Habel. "From Temporal Expressions to Temporal Information: Semantic Tagging of News Messages". *In Proceedings of the ACL 2001 Workshop on Temporal and Spatial Information Processing*, 2001.

[11] Benyah Shaparenko, Rich Caruana, Johannes Gehrke, and Thorsten Joachims. "Identifying Temporal Patterns and Key Players in Document Collections". *In Proc. of the IEEE ICDM Workshop on Temporal Data Mining: Algorithms, Theory and Applications (TDM-05)*, 165–174, 2005.

[12] SIMILE Timeline toolkit, http://simile.mit.edu/timeline/

[13] TimeML, Markup Language for Temporal and Event Expressions, http://www.timeml.org/

[14] TimeWall, http://www.inxight.com/products/sdks/tw