TimeTrails: A System for Exploring Spatio-Temporal Information in Documents

Jannik Strötgen
Institute of Computer Science
University of Heidelberg
Heidelberg, Germany
stroetgen@informatik.uni-heidelberg.de

Michael Gertz
Institute of Computer Science
University of Heidelberg
Heidelberg, Germany
gertz@informatik.uni-heidelberg.de

ABSTRACT

Spatial and temporal data have become ubiquitous in many application domains such as the Geosciences or life sciences. Sophisticated database management systems are employed to manage such structured data. However, an important source of spatio-temporal information that has not been fully utilized are unstructured text documents. In documents, combinations of temporal and spatial expressions form events, which can be mapped to a database structure and organized into trajectories that can be explored. In this context, the coupling of information retrieval techniques with spatio-temporal database concepts leads to new ways for managing and exploring document collections.

In this demonstration, we present TimeTrails, a system for the extraction, querying, storage, and exploration of spatio-temporal information embedded in text documents. The user can query a document collection, and TimeTrails visualizes the spatio-temporal information extracted from relevant documents as document trajectories, resulting in a map-based view of documents. This view helps the user to explore the temporal and spatial content of documents in a meaningful way and to further restrict search results using spatial and temporal predicates.

1. INTRODUCTION

Driven by major advancements in sensor technology and instrumentation of experiments, spatio-temporal data have become ubiquitous in disciplines such as the Geosciences, physical sciences and life sciences. Sophisticated database management infrastructures are employed that allow to efficiently store, query and analyze spatio-temporal data. However, there is also a huge amount of spatio-temporal information embedded in many textual documents, such as historical texts, collections such as Wikipedia, and biographies, to name only a few. In documents, spatio-temporal information naturally corresponds to the textual description of events that take place at some geographic location at some point or duration in time. While there have been advances in the IR community to utilize temporal and geographic information embedded in documents (see, e.g., [1, 7]), these approaches have primarily been developed in isolation, either solely focusing on temporal information or only on geographic information in documents. Only recently, there have been attempts to combine these approaches [4], and a general model for spatio-temporal information in documents has been described [9]. Given the functionality provided by spatio-temporal data management systems, it is desirable to use these techniques also for querying and analyzing spatio-temporal information embedded in documents.

In this demonstration, we will show how a composition of information retrieval and database management techniques can be employed to support new ways for exploring document collections with a spatio-temporal focus. Our system, called TimeTrails, enables the extraction of temporal and spatial information, typically represented in the form of textual expressions, from documents and the management of such information in a database system. A key novel feature of our system is that spatio-temporal information extracted from documents is not just considered in isolation, but relationships between such "events" are derived, resulting in descriptions of event sequences very similar to object trajectories studied in the context of moving object databases. Having the spatio-temporal information from documents readily available in a database system provides us with effective means to support various visualization, query, and exploration scenarios for document collections.

In the following section, we will give an overview of the core concepts and components underlying our system. In Section 3, we then detail some key scenarios and use cases that will be demonstrated using TimeTrails.

2. SYSTEM ARCHITECTURE

TimeTrails is a system for the extraction and exploration of spatio-temporal information embedded in text documents. The system is motivated by (1) the assumption that this information is characteristic for many documents that textually describe events and (2) the need for better ways to explore documents beyond just browsing or applying standard information retrieval techniques. To combine spatial information with temporal information extracted from documents, a model is needed that precisely defines spatial and temporal information in documents (or rather the corresponding textual expressions) and how to combine such information. Our model uses co-occurrence information about

Proceedings of the VLDB Endowment, Vol. 3, No. 2

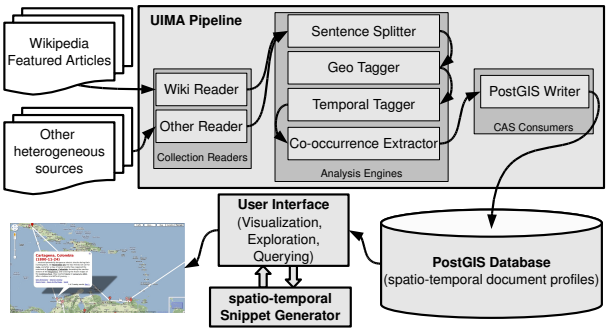

Figure 1: System architecture of TimeTrails with its three components: UIMA pipeline, database, and user interface.

spatial and temporal expressions detected in documents. For this, a window is used in which the expressions have to co-occur to form an "event". For example, a temporal expression (such as "May 15, 2009") that co-occurs with a spatial expression (such as "Berlin") in the same sentence is identified and recorded as spatio-temporal information. For the extraction of sentences, temporal and spatial expressions, and for identifying co-occurrences, a *document processing pipeline* is used. The pipeline is one of three components of the TimeTrails system. The other two components are (1) a PostGIS database for managing and querying documents and extracted information and (2) a user interface component for querying, exploring and visualizing spatio-temporal information about documents on a map. The complete architecture of TimeTrails is shown in Figure 1.

2.1 Document Processing Pipeline

The document processing pipeline used for the extraction of spatio-temporal information from documents is based on the Unstructured Information Management Architecture (UIMA) [10], which is widely used for processing unstructured content like text, audio, or images. All components of a UIMA pipeline use the same data structure, the *Common Analysis Structure* (CAS), resulting in an easy way to combine tools that were not originally built to be used together. This has the advantage that existing tools or new components can be integrated easily. In general, a UIMA pipeline consists of three types of components. First, a *Collection Reader* is used for accessing the textual documents from a source (file system, Web, etc.) and initializing a CAS object for every subject of analysis. Then, the *Analysis Engines (AEs)* perform an analysis of the documents, extract information and add annotations to the CAS objects. Finally, a *CAS Consumer* is used for the final processing of the documents, e.g., to store the annotated information in a database.

As shown in Figure 1, the document processing pipeline consists of several Collection Readers for accessing different types of document collections. A collection reader we use in our demonstration is the WikiReader, which accesses the corpus of Wikipedia Featured Articles [11]. An example of another type of document collection are hit lists returned by traditional search engines, thus allowing our system to be plugged into a search engine.

For the extraction of co-occurrences of spatial and temporal expressions, Analysis Engines are used for splitting a document into sentences, and for extracting and normalizing spatial and temporal expressions. For splitting a document into sentences, we use the OpenNLP Sentence Splitter [6]. Temporal tagging and geo-tagging are typical *named entity recognition* tasks. Temporal expressions are extracted using HeidelTime [8], which is mainly based on regular expressions for the identification of temporal expressions and uses linguistic clues for the normalization of the expressions. HeidelTime was the best-performing system for the extraction and normalization of temporal expressions from English documents in the TempEval-2 challenge. Geolocations (or geospatial entities) are identified using the MetaCarta GeoTagger API [5], which is based on a gazetteer. The final Analysis Engine uses the results of the former AEs, i.e., the information about sentences as well as spatial and temporal expressions, to identify co-occurrence patterns of spatial and temporal expressions.

As the last component of the pipeline, a CAS Consumer is used for storing the extracted information in the form of so-called spatio-temporal document profiles, which are described in the following section. The extraction tasks of the document processing pipeline are detailed in [9]. In our demonstration, we focus on the TimeTrails system itself, in particular the functionality for the visualization and exploration of the information extracted and recorded from documents using the processing pipeline described above.

2.2 Spatio-Temporal Document Profiles

For every document, the extracted spatial and temporal information is managed in a PostGIS database. For this, so-called *spatio-temporal document profiles* are used [9]. In general, a document profile can be considered a complex object that precisely describes a list of time/location pairs that have been identified in a document.

More specifically, a spatio-temporal document profile for a document d, denoted $stdp(d)$, is a sequence of ordered tuples $\langle t, s\rangle$, with t specifying a temporal expression (e.g., "April 15, 2010"), its normalized value (e.g., "2010-04-15"), and its offset position in d. The component s specifies a spatial expression, its grounded value as $\langle$latitude, longitude$\rangle$ pair, and its document offset. For $\langle t, s\rangle$ being an element in $stdp(d)$, the expressions described by t and s have to co-occur in the same sentence E of the document d. This is determined using the offset information, i.e., the start and end positions of the sentence E, the temporal expression t, and the spatial expression s. E, t, and s are extracted using the document processing pipeline described in Section 2.1.

The normalization of temporal and spatial expressions allows TimeTrails to manage such information in a structured format in PostGIS. More precisely, the spatial data in s are indexed using an R-Tree to efficiently support region and nearest-neighbor queries, and SQL (with its extension to support querying spatial and temporal data) is used to efficiently support querying and exploring documents.

The ordering of tuples in $stdp(d)$ is done chronologically based on the normalized time values (chronons). If two tuples contain the same temporal value in their t component, a heuristics based on document order is used for deciding which one is listed before the other. An example of a short part of a spatio-temporal document profile is given in Table 1. The ordering of time/location pairs based on the temporal component naturally leads to a so-called *document trajectory*, similar to trajectories known from moving object

	temporal information			spatial information			
	normalized value	expression	offset	latitude/longitude	expression	offset	
stdp(d) = {...,							
⟨	1882-02-02,	"February 2, 1882",	2882-2898,	-6.26833/53.3122,	"Rathgar",	2952-2958	⟩,
⟨	1887,	"1887",	3235-3239,	-6.25/53.33,	"Dublin",	3432-3438	⟩,
⟨	1931,	"1931",	15836-15840,	-0.1/51.52,	"London",	15826-15832	⟩, ...}

Table 1: Some entries of the spatio-temporal document profile of the Wikipedia article *James Joyce*.

databases, an aspect that provides interesting exploration functionality for documents, as we will show in Section 3.

2.3 User Interface

The third module of TimeTrails is the user interface, which allows the user to search for (relevant) documents and which realizes the visualization and exploration of the spatio-temporal information associated with (selected) documents.

Using a Web-based interface, the user can query a document collection. From the hit list returned by the underlying search engine (in TimeTrails, we use Lucene [3]), the user then can choose one or more documents for which the respective document trajectories are to be visualized on a map. An example of the visualization of the trajectories for multiple documents is shown in Figure 2. For the visualization, TimeTrails uses the Google Maps API [2]. After the user selects documents from a hit list, the corresponding spatio-temporal document profiles are retrieved from the database and used for generating a KML file, which is then visualized using the Google Maps API.

For each location in a document trajectory, furthermore a spatio-temporal snippet is managed in the database. Snippets are calculated using the spatio-temporal Snippet Generator integrated in TimeTrails. The original document and the offset information of the time/location pair are used to create a snippet, which shows a fixed number of tokens before and after these expressions. The minimum size of a snippet is always a complete sentence in which the spatial expression co-occurs with the temporal expression. For better readability, all spatial and temporal expressions in spatio-temporal snippets are displayed in either italics or are underlined, as shown in Figure 2.

3. DEMONSTRATION

In our demonstration of the TimeTrails system, we will focus on the querying, exploration, and visualization aspects of spatio-temporal information embedded in and extracted from documents. TimeTrails provides two Web-based search and exploration interfaces, one corresponding to a traditional text search engine (in our case the full text search library of Lucene) and one corresponding to a map as a visual representation of an area.

In the scenarios described in the following, we use the Wikipedia Featured Articles [11] as corpus. These articles are determined by the editors to be the best articles in Wikipedia. In addition to this subset that contains more than 2800 articles out of 29 categories, we added about 200 articles dealing with history or biographies, which are two subjects that fit very well in our spatio-temporal document exploration framework.

In the following Section 3.1, we detail the query and visualization scenarios, and in Section 3.2, we detail some interesting search and exploration scenarios based on a map.

3.1 Visualization

Using the Web-based search interface, the user can enter a text query, and a hit list of documents relevant to the search query is displayed. With each document in a hit list, the number of identified time/location pairs is shown as well as the range and number of temporal expressions in the form of a sparkline. This information already gives the user a good idea of the amount of spatio-temporal information identified for each document, as not for all documents and document collections a spatio-temporal exploration is meaningful. Based on a hit list, the following scenarios will be demonstrated.

Single Document Visualization (SDV). In the first scenario, the user selects a document from the hit list for a map-based visualization. TimeTrails visualizes the trajectory of this document on the map, i.e., all geographic locations that occur in the spatio-temporal document profile of the selected document are shown on the map, connected by directed lines, representing the document trajectory, i.e., the temporal relations between the individual locations.

Multiple Document Visualization (MDV). Here, the user can select multiple documents from the hit list. The MDV-view displays the document trajectories of the selected documents at once. The trajectories are visualized in the same way as in the SDV-view. The user can choose to add or delete trajectories corresponding to documents of the hit list. An example of the visualization of two document trajectories is shown in Figure 2.

Spatio-Temporal Document Snippets. In both, the SDV- and MDV-view, the user can click on a location associated with a document trajectory to display a spatio-temporal snippet, showing the part of the document text in which the co-occurring expressions (both are highlighted) corresponding to time/location pair have been identified. The user can also select a directed line between two locations and the two snippets associated with the start and end location of that line are displayed at once. This gives the user an idea of the duration between the two events at the two locations, again very much like in traditional object trajectories. In particular, these snippets allow the user to explore specific spatio-temporal information without having to go through the whole document.

Intersecting Document Trajectories. An interesting scenario for exploration is when the trajectories of two documents intersect at a location that has been identified in both documents. Here, we distinguish two cases. First, the location and corresponding expression occur in the two documents but have different times associated (i.e., the two temporal expressions are disjoint and their normalized time values/intervals do not intersect). In such a case, two "events" happened at that location, but at different times. Second, the two events at that location can happen at the same

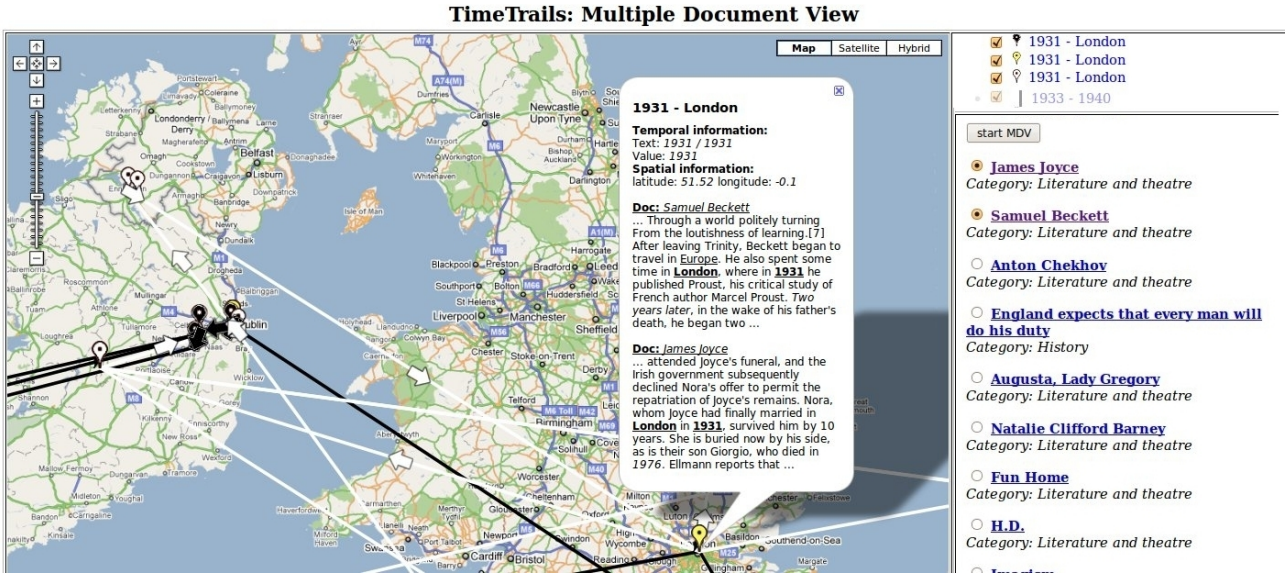

Figure 2: Snapshot of TimeTrails showing the Google Maps view, containing parts of two document trajectories of the Wikipedia articles *Samuel Beckett* (white) and *James Joyce* (black).

time, meaning that the two corresponding normalized time expressions intersect. This situation is indicative for a scenario where the two documents (partially) have the same spatio-temporal scope. For example, given two biography documents, such a scenario can occur when both documents mention that the two people met at some place and time. Technically, for a given time/location pair in a document trajectory, a spatio-temporal join based on other documents' profiles is used to determine such cases. On the map, such intersections are visualized as a yellow pin, as depicted in Figure 2 for the location "London".

3.2 Map-based Querying and Exploration

In addition to the standard text search interface for a document collection described above, TimeTrails also provides the user with a map-based query interface (the same used for visualizing document trajectories). Search and exploration scenarios in this interface focus on spatial and temporal properties, as outlined below.

Region Queries. In this search scenario, the user specifies a rectangular region on the map. A list of documents is determined such that each document is "relevant" to that region. For this, at least one location specified in a document profile must be in that region. Documents are ranked depending on how many of such locations are specified in the profile. To test whether a location (as latitude/longitude pair) is contained in the region, the spatial index is used. For each such document, the user then can choose to visualize the document's trajectory.

Temporal Restrictions. In addition to a region query, the user can also specify a time interval. In this case, only documents relevant to both the query region and time interval are determined and ranked. Again, document profiles, as a structured representation of spatio-temporal information associated with each document, are used to determine and rank the documents. Respective documents are shown in a hit list, and the user can choose to display the trajectories for the documents on the map.

Spatio-Temporal Document Similarity. The trajectories in an MDV-view can be used to explore different documents with respect to their (partial) spatio-temporal similarity. Once two document trajectories overlap at some locations, the user directly sees that both documents contain similar spatio-temporal information and thus are partially similar in terms of their spatio-temporal content.

4. REFERENCES

[1] O. Alonso, M. Gertz, and R. Baeza-Yates. On the value of temporal information in information retrieval. *SIGIR Forum*, 41(2):35–41, 2007.

[2] Google Maps API. `http://code.google.com/apis/maps/`.

[3] Lucene. `http://lucene.apache.org/`.

[4] B. Martins, H. Manguinhas, and J. Borbinha. Extracting and exploring the geo-temporal semantics of textual resources. In *Proc. of the International Conference on Semantic Computing*, 2008.

[5] MetaCarta Inc. `http://www.metacarta.com/`.

[6] OpenNLP. `http://opennlp.sourceforge.net`.

[7] R. Purves, P. Clough, and C. Jones, editors. *Proceedings of the 6th Workshop on Geographic Information Retrieval*, 2010.

[8] J. Strötgen and M. Gertz. HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions. In *Proc. of the 5th International Workshop on Semantic Evaluations*, 2010.

[9] J. Strötgen, M. Gertz, and P. Popov. Extraction and exploration of spatio-temporal information in documents. In *Proc. of the 6th Workshop on Geographic Information Retrieval*, 2010.

[10] UIMA. `http://uima.apache.org/`.

[11] Wikipedia Featured Articles. `http://en.wikipedia.org/wiki/wikipedia:Featured_articles`.