Extraction and Exploration of Spatio-Temporal Information in Documents

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

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

Pavel Popov
Inst. of Computer Science
University of Heidelberg
Heidelberg, Germany
pavel.popov@urz.uni-hd.de

ABSTRACT

In the past couple of years, there have been significant advances in the areas of temporal information retrieval (TIR) and geographic information retrieval (GIR), each focusing on extracting and utilizing temporal and geographic information, respectively, from documents for search and exploration tasks. Interestingly, there is only little work that combines models, techniques and applications from these two areas to support scenarios and applications where temporal and geographic information in combination provide interesting meaningful nuggets in document exploration tasks, such as visualizing a chronological sequence of events with their locations.

In this paper, we present an approach that combines the two areas of TIR and GIR. Using temporal and geographic information extracted from documents and recorded in temporal and geographic document profiles, we show how co-occurrences of such information are determined and *spatio-temporal document profiles* are computed. Such profiles then provide the basis for a variety of document search and exploration tasks, such as visualizing the sequences of events on a map. We present a prototypical implementation of our system and demonstrate the effectiveness of combining GIR and TIR in the context of document exploration tasks.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Linguistic processing; I.2.7 [**Natural Language Processing**]: Text analysis

General Terms

Theory, Languages

Keywords

Information retrieval, temporal data, spatial data, text mining, UIMA

1. INTRODUCTION

In addition to traditional information retrieval capabilities supported by today's search engines, in the past couple of years, more and more search and exploration tools have emerged that focus on detecting and exploiting different types of so-called named entities in documents. Such entities typically correspond to places or locations, events, organizations or people [22]. Example named entity recognition (NER) tools include systems such as ANNIE, Inxight, or OpenCalais. Aspects related to temporal and geographic information embedded in documents that go beyond just the timestamp and location of publication of a document have been of particular interest in many practically relevant document search and exploration tasks. Almost all types of documents contain a variety of temporal and geographic information that describes events, the location of such events, typically in combination with other named entities such as persons or organizations.

Geographic Information Retrieval (GIR) typically has focused on extracting geospatial information from documents and utilizing such information to filter documents or to provide a geographic focus of a document (e.g., a document primarily talking about locations in Japan) [15]. On the other hand, Temporal Information Retrieval (TIR) concerns the extraction of temporal information from documents, such as explicit temporal expressions (dates) or relationships between events. Such information, too, provides important insights into documents, such as the timespan covered in a text or the temporal focus of a document (see, e.g., [3]).

Although individual TIR and GIR methods, models and tools lead to important improvements of traditional search engines, as now documents can be filtered based on some geographic or temporal properties (e.g., news or RSS feeds), interestingly, there is little work that combines these two areas. The motivation of our work is the fact that events typically happen at some location. That is, oftentimes, temporal information can be associated with a location found in some document and vice versa. Combining respective techniques from TIR and GIR then would allow, for example, to filter documents on both geographic and temporal properties. However, such a combination may even lead to more advanced document search and exploration methods and tools. Our motivating example for this is the discovery of event/location sequences from documents, as they can be found in travel reports and documents related to history.

To support such tasks, in this paper, we present a model and tool that facilitates exploiting temporal and geographic information that has been extracted from documents using traditional NER tools. Our particular focus is on document content where geographic and temporal expression co-occur, as an indication for some information about an event/location pair. We detail how temporal and geospatial information extracted from documents is described in temporal and geographic document profiles. These two types of

profiles, having a TIR and GIR focus, respectively, are then combined in so-called spatio-temporal document profiles that specify event/location pairs in a chronological order.

While temporal and geographic profiles alone are already useful for different document search and exploration tasks, their combination allows to describe some kind of document trajectories, that is, a chronological enumeration of locations related to temporal expressions. Especially such trajectories lead to an interesting visualization aspect in support of document exploration. Another contribution of our work is the description of a complete processing pipeline that combines TIR and GIR techniques to determine spatio-temporal document profiles.

The remainder of the paper is structured as follows. After a review of existing work on TIR and GIR in the following section, in Section 3, we present the concepts underlying document profiles that focus on temporal expressions, geographic expression and combinations thereof. In Section 4, we present our prototypical implementation of a system to determine so-called spatio-temporal document profiles from documents, with a particular focus on the extraction of temporal information. In Section 5, we present some advanced search and exploration tasks that utilize spatio-temporal information extracted from documents. We conclude the paper with a summary and outline of ongoing work in Section 6.

2. RELATED WORK

A basic prerequisite underlying our model and approaches to exploring spatio-temporal information in documents are tools and techniques for the extraction of temporal and geographic information from documents. In the following, we first give an overview of Named Entity Recognition (NER) techniques related to temporal information extraction and then focus on geographic information extraction. In both cases, we also outline existing approaches to using such information for document search and exploration tasks. Finally, we give an overview of approaches that combine extracted temporal and geographic information for more advanced document search tasks.

2.1 Temporal Information

Research on temporal information extraction concerns the identification of temporal expressions and relations between such expressions. One way of annotating identified temporal expressions is to use the TimeML markup scheme, which is an ISO standard for the annotation of temporal expressions, events and relations between them [26]. TimeBank is a corpus that consists of 182 news documents that are manually annotated according to the TimeML specification [27]. An analysis of the corpus is given in [4]. A temporal tagger for the extraction and normalization of temporal expressions is GUTime [8], which uses the TimeML Timex3 standard for annotating temporal expressions. GUTime extents the TempEx tagger, which used the ACE Timex2 standard for the recognition and normalization of temporal expressions [17]. GUTime is one component of the Tarsqi toolkit, which includes other components for the extraction of events and relations between temporal expressions and events [32]. The TimeBank corpus was also used for the TempEval-1 challenge, where participants had to develop techniques to extract temporal relations between events and between events and temporal expressions. The annotations of both events and temporal expressions are given so that only the temporal relations had to be extracted [31].

In general, the central role of time in any information space has been studied not only in the area of information extraction, but also in other areas related to information retrieval, such as question answering and summarization [16]. A discussion of different document search and exploration tasks focusing on temporal information embedded in documents is given in [2]. Temporal information extracted from documents can be used for a variety of tasks, such as constructing timelines for clustering and exploring document search results [3] or associating temporal snippets with documents in a hit list [1]. Underlying these techniques are often temporal document profiles, which are constructed by NER tools to represent extracted temporal information from documents for further processing and analysis.

2.2 Geographic Information

As for temporal information extraction, there is a lot of research on extracting geographic information from documents and using such information in different search tasks. For the extraction and normalization of geographic information, typically a NER tool utilizes a gazetteer. For the extraction of named entities in general, gazetteers do not necessarily improve extraction results, but for the extraction of location names, gazetteers are extremely valuable [21]. When named entity resolution is needed, i.e., when extracted geo-locations need to be associated with spatial information such as latitude and longitude, a gazetteer is essential for finding all alternatives of a location [12]. For disambiguation among alternatives, the context in which the geographic expression occurs is often taken into account by applying linguistic rules and heuristics, for example, that geographic references in the same discourse are somehow (spatially) related to each other [13]. A detailed survey of techniques for place reference disambiguation is given in [11]. There are a few publically available tools and services for the extraction of named geographic entities, such as MetaCarta [20] and OpenCalais [23].

There are several applications for which extracted and normalized spatial information is crucial. Lieberman et al. describe a spatial textual search engine called STEWARD, which processes unstructured documents, extracts spatial information and adds a geographic focus to the documents, which is then used for visualizing the documents on a map [14]. Google Books offers the ability to visualize all locations mentioned in a book and Google News can be used to cluster news documents by location.

The GeoCLEF tracks of the Cross Language Evaluation Forum focus on the evaluation of IR systems with respect to geographic relevance [15]. Most of the documents and queries considered in this context contain some kind of geographic information, and the participating systems are evaluated with respect to this information. The Geographic Information Retrieval series of workshops are another example showing the activity of research in this field [9].

2.3 Spatio-Temporal Information

Although there have been substantial developments in each of the two areas, temporal information extraction and geographic information extraction, there is only little work combining both types of extracted information in document search and exploration tasks. One of the first work combining temporal and geographic information is GeoTracker, which reorganizes RSS feeds of news documents according to the RSS feed time and locations mentioned in these feeds [5]. For the extraction of geographic information, a rule-based tagger has been developed to extract explicitly mentioned locations. Extracted entities are matched against a location database for normalization purposes. Finally, generic matches are eliminated if a more specific match is extracted in the same RSS feed. In contrast to our work, GeoTracker associates only one place and one point in time with every document (RSS feed). The RSS feeds are visualized on a map and can be traced over time using a time slider. While combining temporal and geographic informa-

tion, GeoTracker only uses geographic information extraction because the only temporal information used is the time of the RSS feed, i.e., the document creation time and not other temporal expressions that occur in a document.

Approaches that consider both, the extraction of temporal and spatial information, are covered by Gey et al. in [7] and Martins and colleagues in [18, 19]. Although Gey et al. treat biographies as ordered sequences of events in time and space by defining events as a 4-tuple (Activity, Date-range, Place, Other people), they use the temporal information only for defining what events to visualize and do not visualize the temporal information itself [7]. The work by Martins et al. focuses on RSS feeds and extracts temporal and geographic information from such feeds [18]. They extended this system by using similar methods for textual resources, i.e., descriptive metadata in digital libraries [19]. Their focus is on the extraction of temporal and geographic information based on a gazetteer to translate the textual descriptions of time periods and place names given in the metadata of records in the digital libraries. Finally, they add a temporal and a geographic scope to each document. Their temporal information extraction focuses on complete dates and names for historical periods. In contrast to our work, they are not trying to disambiguate incomplete expressions of temporal information.

3. COMBINING TEMPORAL AND SPATIAL DOCUMENT PROFILES

The idea behind so-called spatio-temporal document profiles is to describe all temporal and geographic information extracted from a document in a concise manner before such information is further utilized in search and exploration tasks. Before we detail such profiles, in the following Sections 3.1 and 3.2, we first describe the concepts underlying temporal profiles and spatial profiles, respectively, before we combine such profiles into comprehensive spatio-temporal document profiles in Section 3.3.

3.1 Temporal Document Profiles

As the basis of the temporal document profiles, we assume a discrete representation of time based on the Gregorian Calendar, with a single day being an atomic time interval called *chronon*. Our *base timeline*, denoted T_d, is an interval of consecutive day chronons. For the representation of different granularities, we assume different timelines $\mathcal{T} = \{T_d, T_w, T_m, T_y\}$ for day, week, month, and year, respectively. The value of a finer granularity can be mapped to a coarser granularity since months and weeks consist of days, and years consist of months. In addition, we introduce a *precedence relationship* $\prec_T$ that allows to compare chronons. For two chronons $t_i, t_j \in \mathcal{T}, t_i \neq t_j$, either $t_i \prec_T t_j$ or $t_j \prec_T t_i$.

For being able to build up temporal document profiles and to use the assumptions of timelines and of the precedence relationship, a key component to our approach is to extract all types of temporal information associated with a document. The first type of such information is the *document timestamp*, which appears as the date a document has been created or last modified. This point in time can be anchored in the timeline T_d. The next type of temporal information are *temporal expressions*, which can be distinguished between explicit, implicit, and relative temporal expressions [28]. Explicit temporal expressions describe chronons in some timeline, such as "May 2009" that can be anchored in T_m or "March 11, 2006" that can be anchored in T_d. *Implicit temporal expressions*, such as names of holidays or events, can be anchored in a timeline. For example, "Columbus Day 2006" can be mapped to the expression "October 12, 2006", which is anchored in T_d. *Relative temporal expressions* can only be anchored in a timeline in reference to another already anchored temporal expression. For example, the expression "next month" alone cannot be anchored in any timeline. However, it can be anchored if the document is known to have a creation date as reference.

Based on the assumption that temporal expressions are determined by a temporal tagger, we define a temporal document profile, denoted *tdp(d)*, as consisting of the extracted temporal expressions. A temporal document profile thus is a sequence of tuples $\langle e_i, c_i, p_i \rangle$ with e_i being the temporal expression, c_i an element from the timelines $\mathcal{T} = \{T_d, T_w, T_m, T_y\}$ corresponding to the expression e_i, and p_i describing the offset information of the expression in the document. An offset specifies the start and end position of an expression in a document. All chronons c_i in the temporal document profile are normalized to their ISO format for temporal expressions (see Section 2.1). For instance, all day chronons are represented in the year-month-day format, such as "1982-03-11".

3.2 Spatial Document Profiles

Analogous to temporal document profiles, a spatial document profile describes the geographic information extracted from a document. Such information typically describes some geographic *locations* that are of different types of granularity and spatial extent, respectively (e.g., streets, cities, countries).

There are several components associated with extracted location information. First, a location should be normalized. That is, different expressions in a document may refer to the same geographic entity. For example, the geographic expressions "US", "U.S.A.", and "United States" all correspond to the (normalized) expression and entity "United States of America". Such normalization, as part of a geo-tagger, is rarely available for all types of locations but would simplify associating spatial information with a location such as position or spatial extent.

The second component associates a *geometry* with a location expression. A geometry is either a point or a polygonal region. We assume latitude/longitude pairs to describe such geometries. For example, if a name of a city has been identified as a geographic expression in a document, the associated geometry can be either a point or a polygon describing the boundary of the city, the latter corresponding to some typical vector data. Associating geometry information with geographic information identified in a document, of course, heavily depends on how comprehensive and detailed the gazetteer used by the underlying NER tool is. The MetaCarta service mentioned in Section 2.2, for example, provides only point-based geometries for a location, no matter what type the location is. That is, for both an address and a city, a point geometry is determined. However, with MetaCarta, the same geometry information is associated with different expressions if these expressions refer to the same geographic entity. In the case normalization of geographic expressions is available, e.g., based on the location names used in OpenStreetMap [25], then region-based geometries could be associated with locations, because OpenStreetMap also manages vector data associated with entities such as cities, countries, and other administrative units in many countries. In the following, as part of our conceptual model, we assume that only with geographic locations of the type address point-based geometries are associated. For more complex location types such as cities and countries region-based geometries are assumed.

Geometries naturally induce some topological relationships such as containment or overlap. In our approach, we are particularly interested in containment relationships between locations and their corresponding geometries, because this leads to a natural hierarchy among location types. For example, a city (with a region geometry) is contained in a country (also with a region geometry), and an

address (as point geometry) is contained in both. In the following, if the geometry of an expression g_i is contained in the geometry of an expression g_j, we denote this with $g_i \subset g_j$. This reflects an intuitive view on geographic information contained in a document.

Based on the above discussion, for a given document d, we assume a *spatial document profile,* denoted *sdp(d)*, is associated with d after a geographic information extraction process. The profile consists of a set of tuples $\langle g_i, v_i, p_i \rangle$, where (1) g_i is the geographic expression as it occurs in d, (2) v_i represents the geometry that has been determined for the location corresponding to the expression g_i, and (3) p_i is the offset information of the expression g_i in d.

As one can easily see, the difference to a temporal document profile is that a geometry is associated with a geographic expression instead of a normalized chronon for a temporal expression.

3.3 Spatio-Temporal Document Profiles

The objective of spatio-temporal document profiles is to combine the information represented in temporal and geographic profiles in a meaningful way. In our approach, for a given document d, we are in particular interested in pairs of tuples $\langle t, s \rangle$, $t \in tpd(d)$ and $s \in sdp(d)$, for which the temporal expression occurs close to the geographic expression. This is intuitive as events (represented by a temporal expression) typically happen somewhere, i.e., at a location. Although there are many alternatives of how "occurs close to" can be defined, we define a geographic expression co-occurring with a temporal expression, if both expressions occur within a specific window size in the document, for example, within one paragraph or one sentence. This can easily be determined based on the offset recorded for each tuple t and s. The offset information of the expressions is then compared to the offset information of the window. An advantage of using co-occurrences is that almost no preprocessing is necessary, thus resulting in a very fast processing time for determining such co-occurrences. A disadvantage of this approach is that hardly any context information is taken into account. For example, if there is more than one temporal expression and more than one geographic expression within one sentence, there is no easy way to decide which temporal expression should be associated with which geographic expression. In such cases, we use the cross product to enumerate pairs of respective tuples.

For a document d, we define a spatio-temporal document profile, denoted $stdp(d)$, as an *ordered* list of tuples as follows. Given two tuples $t = \langle e, c, p_t \rangle$ and $s = \langle g, v, p_s \rangle$ such that the expressions e and g co-occur in a sentence. Then the tuple $st = \langle e, c, g, v, p_t, p_s \rangle$ is an element of $stdp(d)$. The tuples in $stdp(d)$ are sorted by the chronons c_i, beginning with the earliest chronon and ending with the most recent chronon (or the most distant in the future).

Sorting tuples in this way can be applied, because with every temporal expression a normalized chronon is associated. However, two problems may occur. First, several tuples in $stdp(d)$ can have the same normalized chronon (i.e., $c_i = c_j$ for two tuples $st_i, st_j \in stdp(d)$). Second, the chronons of two tuples can be of different granularities ($g(c_i) \neq g(c_j)$, with $g(c) \in \{\text{year, month, week, day}\}$ for any chronon c). For example, the normalized chronons "2009-03" and "2009-03-10" are not comparable in terms of the precedence relationship $\prec_T$.

The first problem is solved by considering in which sentence (and thus at what position in the document) the expressions associated with a tuple occur. For this, we assume that if the expressions e_i, g_i occur before the expressions e_j, g_j in the document, then tuple st_i comes before tuple st_j in $stdp(d)$. This reflects the intuition that events (and their locations) are mentioned in some chronological order in a document. The second problem of non-comparable chronons is solved as follows. Given two chronons c_i and c_j that are not comparable in terms of $\prec_T$, e.g., "2009-03" and "2009-03-10". Then, the chronon, with the finer granularity is converted into a chronon with the granularity of the coarser chronon. For the two chronons above, we then obtain c_i = "2009-03 and c'_j = "2009-03". Now these two chronons can be compared using the method discussed in the previous paragraph where the position of the expression is used.

Using the above methods, the tuples in a spatio-temporal document profile can be totally ordered. Such a profile now can be used to explore so-called document trajectories, which are discussed next.

A spatio-temporal document profile can be viewed as a description of event/location pairs that are presented in a chronological order. This view resembles the concept of trajectories used in moving object databases, where an object trajectory is a sequence of time/location pairs [29]. We adopt this view in the following sections in which the information specified in a spatio-temporal document profile is used to visualize and explore sequences of time/location pairs determined for a document, for example, to convey a sequence of events in a purely geographic view.

4. DOCUMENT PROCESSING PIPELINE AND EVALUATION

In this section, we present our document processing pipeline. One component of this pipeline is the temporal tagger described in Section 4.1, followed by an evaluation. For this, we compare the results of our temporal tagger with existing approaches. Then, in Section 4.2, the text mining pipeline used for the temporal and geographic information extraction tasks and for the detection of geo-temporal co-occurrences is detailed. These co-occurrences are the basis for creating spatial-temporal document profiles.

4.1 Temporal Tagger

The development of our temporal tagger is inspired by GUTime, which was described briefly in Section 2.1. For our purpose, we are mainly interested in temporal expressions that can be anchored in some timeline and chronon, respectively. This does not mean that the temporal expressions themselves have to be explicit, but one has to be able to normalize them to a specific chronon. For example, the expression *a year ago* is not explicit, but the value attribute (chronon) can be determined if an explicit date has already been mentioned in the same document.

For the development of the temporal tagger, we use the TempEval 2007 corpus as a gold standard, which is based on the TimeBank corpus and which is split into a training and an evaluation set. Even though the extraction of temporal expressions was not a task at TempEval 2007, the temporal expressions are annotated in the corpus. They are grouped into four categories, dates, times, durations and sets. As mentioned above, we focus on dates (and the date information of time expressions). For the rule development, we use the training corpus, which consists of 162 documents, having a total of 1.021 temporal expressions of the type *DATE*.

The temporal tagger consists of three stages. First, the extents of all temporal expressions are extracted, i.e., every temporal expression has an associated offset. For this, we create patterns based on regular expressions. Second, the values of all explicit dates are set to their corresponding ISO Format (see Section 2.1) and the values of all expressions that cannot be specified directly are added in an underspecified way. For instance, the value of *a year ago* is set to `UNDEF-last-year` or the value of *July 4* is set to `XXXX-07-04` since the year cannot be detected without context information. Finally, all underspecified values are explored. For the latter example, the year of the value is set to the year of the previously mentioned

(explicit) temporal expression. If this is, for instance, `1776`, the new, fully specified value for *July 4* would be `1776-07-04`.

We evaluate the quality of the extraction of temporal expressions and use the unseen evaluation set of TempEval 2007. Again, we only take the temporal expressions of the type *DATE*. The 20 documents contain 139 such expressions. The results are shown in Tables 1 and 2, together with the results of the training set.

To be able to compare our results to others, the evaluation is carried out under different settings. First, we analyze the extents of temporal expressions and use settings described by Boguraev and Ando in [4]. They use *exact* match and *sloppy* match settings. For the *exact* match setting, the left and right boundaries have to be identical with the gold standard, while for the *sloppy* match setting only the right boundary has to be identical. The *sloppy* variant is introduced due to some inconsistencies in the TimeBank corpus concerning the left boundary items like determiners or predeterminers [4]. In addition to the results of the rule-based system in [4], we add the results of a machine learning system for comparison purposes [10]. The results for both systems that are given in Table 1 are for the complete TimeBank corpus for all types (Date, Time, Duration, and Set) but without classifying the temporal expressions to the correct type. The results of our system are listed separately for the training set and the unseen evaluation set of the TempEval 2007 corpus and are calculated for the temporal expressions of the type *DATE* only. Hence, the results of our system are not completely comparable to the results of the other systems, but nevertheless give a good impression of the performance achieved by our system. Using the *sloppy* measures, we do not achieve the result of [4] due to a lower recall, but for the *exact* matching measures, we outperform [4] and are on one level with [10] although we have the additional difficulty of extracting temporal expressions of the correct type. Both other systems loose performance when the type of the temporal expressions has to be determined [4, 10].

Unfortunately, for both systems, there is no evaluation concerning the normalization of the temporal expressions, i.e., the chronons corresponding to the extracted temporal expressions [4, 10]. Without normalization, extracted temporal expressions are hardly valuable for our purposes, i.e., for creating document trajectories. Therefore, we added two additional evaluation settings for the temporal tagger, *exact+value* and *sloppy+value*, as shown in Table 2. As expected, there is a loss in precision, recall and f-score compared to the settings without the value feature. This can be explained as the normalization adds an additional difficulty. In addition, a true positive for the *sloppy* or *exact* measure results in both, a false negative and a false positive for the *sloppy+value* and *exact+value* measures, i.e., in two errors, if the value feature is not determined correctly. A first error analysis of the normalization showed that many errors occur in the normalization of phrases like *a year ago*, which is sometimes referred to a year or a quarter or an explicit day in the gold standard. Fortunately, these errors do not have an impact on computing (spatio-)temporal document profiles for documents.

Table 1: Evaluation results for the recognition of temporal expressions on the TempEval 2007 training and evaluation sets (*DATE* only). For comparison, the values of a rule-based system [4] and a machine learning system [10] on the TimeBank corpus are given (all types).

	Precision	Recall	F-Score
Timebank (all types)			
[4] sloppy	85.2	95.2	89.6
[4] exact	77.6	86.1	81.7
[10] exact	86.6	79.6	82.8
Training Set (*DATE*)			
own approach sloppy	88.3	86.3	87.3
own approach exact	83.6	81.7	82.6
Evaluation Set (*DATE*)			
own approach sloppy	90.1	84.9	87.4
own approach exact	86.3	81.3	83.7

Table 2: Evaluation results for the normalization of temporal expressions on the TempEval 2007 training and evaluation sets.

	Precision	Recall	F-Score
Training Set (*DATE*)			
sloppy+value	77.4	75.6	76.5
exact+value	74.0	72.4	73.2
Evaluation Set (*DATE*)			
sloppy+value	80.2	75.5	77.8
exact+value	77.9	73.4	75.6

4.2 Text Mining Pipeline

Our text mining pipeline is based on the Unstructured Information Management Architecture, UIMA [30], which allows to process unstructured content of any type (for example text, audio or images) [6]. In addition, UIMA helps to combine different components, which were originally not built to interact with each other since all components use the same data structure, the Common Analysis Structure (CAS).

In general, three different types of components are included in a UIMA pipeline: Collection Readers, Analysis Engines and CAS Consumers. A Collection Reader reads data from a source (file system, database, etc.), decides how to iterate over the documents and initializes a CAS object for every subject of analysis. The Collection Reader may add metadata to the CAS object as well as the document text, which is the most important part for our purposes, since we are processing text documents. The Analysis Engines are the components of a pipeline that analyze a document, find information, and annotate this information in the CAS object. The first Analysis Engine of a pipeline gets the CAS object from the Collection Reader, the other Analysis Engines from the former Analysis Engine. The extracted or derived information, the Analysis Results, typically contain metadata to the content of a document. For example, a sentence splitter adds sentence boundary information to the CAS. The Analysis Engines can access all the information available in the CAS object, namely the document text and the Analysis Results of former Analysis Engines. The last components of a pipeline are the CAS Consumers. They do not add any information to the CAS object, but do the final processing. Possible tasks are, for example, visualization of relevant information, evaluation based on a gold standard, or building up a search index.

In our case, the main reasons for using UIMA are (1) to build a corpus-independent processing pipeline and (2) to be able to add NLP components without having trouble to integrate them. As shown in Figure 1, we developed two collection readers, the TempEval reader and the MySQL reader. The former is used for accessing the TempEval 2007 corpus to evaluate the temporal tagger, while the latter is used to read data from a MySQL database, in which, as in our case, Wikipedia featured articles are stored. All analysis engines can then be used in the same way, no matter which corpus is chosen. New corpora can be processed easily by developing a collection reader that handles format details of the new corpus and makes available the textual information of the new corpus that can then be added as the document text to the CAS object.

In Figure 1, two workflows are shown. On the one hand, the

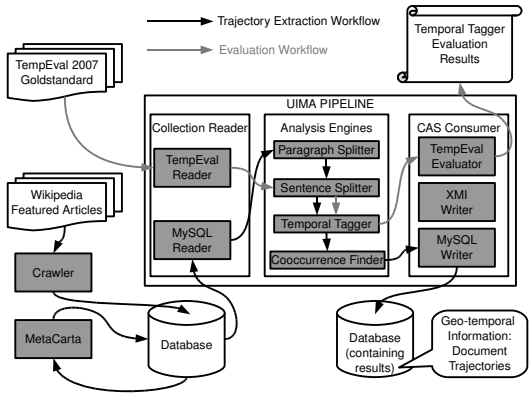

Figure 1: The UIMA-based text mining pipeline with the workflow for evaluation of the temporal tagger and the workflow for document trajectory extraction.

Evaluation workflow we used for evaluating the temporal tagger as described in Section 4.1 and, on the other hand, the Trajectory Extraction workflow we used for our experiments.

As a first step, we crawl the Wikipedia featured articles [33], which we use as our corpus. This subset of Wikipedia that contains more than 2,500 articles out of 29 categories can be seen as a high quality subset of Wikipedia, because the featured articles are determined by the editors to be the best articles in Wikipedia. We store the featured articles in a MySQL database together with some metadata, for example, the title and the category of a document.

For the extraction of geographic expressions, we use the MetaCarta GeoTagger API [20], which is based on a gazetteer and uses linguistic methods for normalizing the found entities. In addition to the longitude/latitude information and a confidence value (that the identified entity is a geographic entity and is normalized correctly), container information is associated with the entities [20]. If a city is extracted, for example, information about the county, state or country is given as well. For our experiments, we use MetaCarta without any adaptations, and we did not do a comprehensive evaluation of the geographic entities, i.e., we use the geographic information as given since we focus on the extraction of document trajectories and the development of the temporal tagger.

As the next step in the text mining pipeline, we develop a UIMA Collection Reader, the MySQL Reader, for accessing the information of the MySQL database. The featured articles and the geographic entities are written to the CAS object. Then, the CAS object is given to the analysis engines. We are splitting the document text into paragraphs and sentences, using a self-developed Paragraph Splitter and the OpenNLP Sentence Splitter [24].

The next analysis engine is the temporal tagger, which was described in Section 4.1. The tagger adds all identified temporal expressions and their normalizations to the CAS object, which now contains annotations of paragraphs, sentences, geographic entities and temporal expressions. The Co-occurrence Finder, which is the last analysis engine in the pipeline, uses all this information for producing co-occurrence pairs of temporal expressions and geographic expressions on two levels, namely: the paragraph level and the sentence level. These co-occurrences are then used for creating the spatio-temporal document profiles, as described in Section 3.3.

5. TIME- AND SPACE-BASED DOCUMENT EXPLORATION

In Section 5.1, we discuss a time- and space-based analysis of the Wikipedia Featured Articles corpus we used for our experiments. In Section 5.2, we present several ideas how to use the spatio-temporal document profiles and document trajectories for search and exploration tasks and show a prototypical visualization.

Table 3: Spatial and temporal statistics on the Wikipedia Featured Articles Corpus.

	average	min	max	std. deviation
temporal ex.	117	0	1557	102
spatial ex.	95	0	788	103
co-occurrences	60	0	796	74

Table 4: Spatial and temporal statistics on the Wikipedia Featured Articles Corpus grouped by categories.

category	average co-occurrences	average temporal expressions	average spatial expressions
Geography and places	142	189	254
Warfare	105	135	158
History	92	127	145
...	...	...	...
Computing	7	76	10
Physics and astronomy	6	99	9

5.1 Corpus Analysis

We use the Wikipedia Featured Articles Corpus for a first analysis of the spatio-temporal document profiles and the document trajectories. For this, we use the Trajectory Workflow of our text mining pipeline as described in Section 4.2 and calculate for each document the numbers of extracted temporal expressions, spatial expressions[1] and co-occurrences, which reflects the length of the spatio-temporal document profile for the specific document. These statistics are shown in Table 3.

The figures show that there is high potential for a spatio-temporal analysis of the documents since the average number in a document is 117 for temporal expressions, 95 for spatial expressions and 60 for co-occurrences at the sentence level. Due to the large difference of the minimum and maximum number of temporal expressions, spatial expressions and co-occurrences, we further analyze the document and focus on the category of the documents in the corpus. Statistics concerning spatial and temporal expressions and co-occurrences for some categories are shown in Table 4. These statistics indicate that the potential for spatio-temporal analyses of the documents depends on the topic. *Geography and places*, *Warfare*, and *History*, for example, are categories for which a spatio-temporal analysis looks very promising, while documents of the categories *Physics and astronomy* and *Computing* are not predestinated for such a type of analysis. For example, the Wikipedia article *Sequence alignment* of the category *Computing* does not contain any spatial information at all and only three temporal expressions.

5.2 Search and Exploration Tasks

For search and exploration tasks, we create the spatio-temporal document profiles for all documents of the Wikipedia Featured Articles. These profiles and the resulting document trajectories are used for search and exploration tasks, for example, for clustering search results, visualization or snippet generation. The clustering

[1]We set the threshold for the confidence that the expression is a spatial expression to 77%. This value was reasonable according to a first analysis of the spatial expressions identified by MetaCarta.

can be done either along timelines, into geographic areas or into geo-temporal categories. The visualization can be done separately, i.e., the geographic information on a map or the temporal information along a timeline. However, using the spatio-temporal document profiles, this information can be visualized in combination by plotting the document trajectories on a map. In addition, the spatio-temporal document profile can be used for snippet generation since the positions in the document of both, the geographic and the temporal expression is available in the document profiles.

A further advantage of using spatio-temporal document profiles for search and exploration tasks is that both, spatial and temporal information, can be used at different granularities. For instance, when searching for documents, a query could be restricted to a specific location and time. The temporal information of all co-occurrence tuples can then be mapped to the desired timeline $\mathcal{T}$ and the spatial information of all tuples to the desired resolution $\mathcal{R}$.

The visualization component of our prototype creates a map for each document in the document collection. The longitude/latitude information of all spatial entities found by MetaCarta in the document is used for creating a KML (Keyhole Markup Language) file. This KML file is used for the visualization of the spatial information on a map using the Google Maps API. In addition to the longitude/latitude information, the spatio-temporal document profile is used for adding the temporal order of all occurring spatio-temporal tuples and information about the position of the tuple in the document. This information is used for connecting all locations being part of a co-occurrence, i.e., being part of the document trajectory, in the correct order. In addition, a snippet for every location is created in which the corresponding passage of the document is shown. In the case that a location is part of the document trajectory, the temporal and the spatial expressions of the co-occurrence tuple are shown in bold and are underlined. All other temporal and spatial expressions that do not belong to this tuple are underlined.

As a prototypical example of such a visualization, Figure 2 shows a part of the document trajectory of Wikipedia's article *Alexander von Humboldt*. The corresponding sentences of this article that include at least one co-occurrence are listed in Table 5. There are sentences between the listed ones that do not contain any co-occurrences of spatial and temporal expressions. On the one hand, some information is missing, for instance, the Orinoco River (sent. 3) is not part of the trajectory because it is not extracted as a location (the confidence is below the threshold). On the other hand, the temporal reasoning works fine, for example, the relative expression *November 24* (sent. 4) in the snippet is correctly normalized to 1800-11-24 because the value of the last mentioned explicit expression is *1800-02* (February 1800 in sent. 3).

Table 5: Sentences of Wikipedia's article *Alexander von Humboldt* containing co-occurrences relevant for the screenshot (see Figure 2). Locations in bold, Temporal expressions in italics.

#	Sentence with co-occurrence
1	... stopped six days on the island of **Tenerife** to climb Mount Teide, and landed at **Cumaná, Venezuela**, on *July 16*.
2	Returning to **Cumaná**, Humboldt observed, on the night of *November 11−12*, a remarkable meteor shower (the Leonids).
3	He proceeded with Bonpland to **Caracas**; and in *February 1800* they left the coast with the purpose of exploring the course of the Orinoco River.
4	On *November 24*, the two friends set sail for **Cuba**, and after a stay of some months they regained the mainland at **Cartagena, Colombia**.
5	Ascending the swollen stream of the **Magdalena**, and crossing the frozen ridges of the **Cordillera Real**, they reached **Quito** on *January 6, 1802*, after a tedious and difficult journey.

6. CONCLUSIONS AND ONGOING WORK

In this paper, we presented a model for spatio-temporal document profiles (stdp) that is valuable for many search and exploration tasks, like clustering, visualization or snippet generation. An stdp makes accessible information about temporal and spatial information that belong together resulting in a document trajectory. To decide whether space and time information belong together, we used a co-occurrence approach. Different window sizes in which the two entities have to co-occur in a document can be chosen.

For the prototypical realization, we developed a UIMA based text mining pipeline to access different document sources, extract spatial and temporal information from documents, find co-occurrences and store the extracted information in a database. In addition, we presented a new temporal tagger that achieves promising results for temporal expressions of the type *DATE* and is competitive to other existing approaches.

We are currently investigating refinements of spatio-temporal document profiles. Having answered *when* and *where*, the next questions to arise are *who* or *what*. This will result in different trajectories for different objects mentioned in the text in addition to the document trajectory. Besides a NER tool for the recognition of objects (persons, real objects or abstract objects like a disease), we are planning to use NLP methods, for example, for coreference resolution. Further NLP methods are also helpful for getting more precise relations among spatial and temporal expressions than the ones that can be extracted using a simple co-occurrence approach. Since our pipeline is based on UIMA, the integration of additional components can be realized easily by including tools that are available as UIMA components.

A better understanding of the syntax and semantics of co-occurrence sentences could help to improve the spatio-temporal document profiles. For example, in the sentence shown in the spatio-temporal snippet in Figure 2, both locations *Cuba* and *Cartagena, Colombia* are associated with the normalized temporal value 1800-11-24. This is not correct for both. Nevertheless, the locations are included in the correct order in the document trajectory because of the assumption that a previously mentioned location occurs before a later mentioned one when both have the same temporal value.

7. REFERENCES

[1] O. Alonso, R. Baeza-Yates, and M. Gertz. Effectiveness of Temporal Snippets. In *WWW '09*, 2009.

[2] O. Alonso, M. Gertz, and R. Baeza-Yates. On the Value of Temporal Information in Information Retrieval. *SIGIR Forum*, 41(2):35–41, 2007.

[3] O. Alonso, M. Gertz, and R. Baeza-Yates. Clustering and Exploring Search Results Using Timeline Constructions. In *CIKM '09*, 97–106, 2009.

[4] B. Boguraev and R. K. Ando. TimeBank-Driven TimeML Analysis. In *Annotating, Extracting and Reasoning about Time and Events. Dagstuhl Seminar Proceedings*, 2005.

[5] Y.-F. Chen, G. Di Fabbrizio, D. Gibbon, S. Jora, B. Renger, and B. Wei. GeoTracker: Geospatial and Temporal RSS Navigation. In *WWW '07*, 41–50, 2007.

[6] D. Ferrucci and A. Lally. Building an Example Application with the Unstructured Information Management Architecture. *IBM Systems Journal*, 43(3):455–475, 2004.

[7] F. Gey, R. Shaw, R. Larson, and B. Pateman. Biography as Events in Time and Space. In *GIS '08*, 89, 2008.

[8] GUTime. `http://www.timeml.org/site/tarsqi/modules/gutime/index.html`.

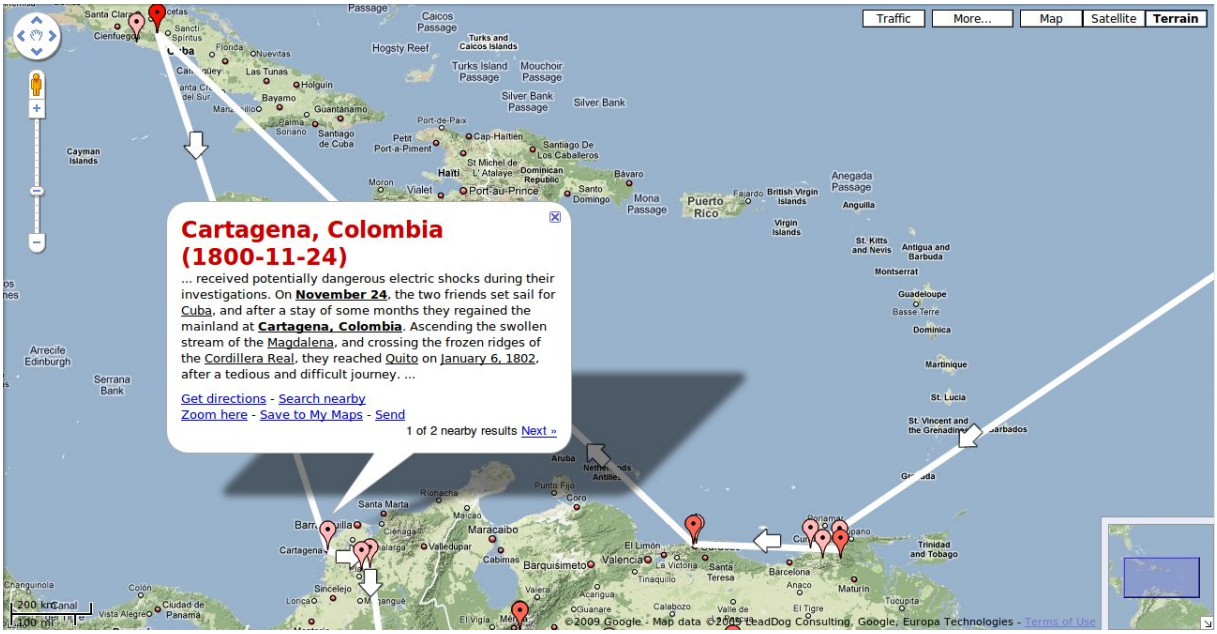

Figure 2: A Google Maps screenshot containing a part of the document trajectory for the Wikipedia Article *Alexander von Humboldt*. All extracted locations are shown. Directed lines between the locations represent the document trajectory.

[9] C. Jones and R. Purves, editors. *Proceedings of the 5th ACM Workshop On Geographic Information Retrieval*, 2008.

[10] O. Kolomiyets and M.-F. Moens. Meeting TempEval-2: Shallow Approach for Temporal Tagger. In *DEW '09: Proc. of the Workshop on Semantic Evaluations*, 52–57, 2009.

[11] J. Leidner. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. PhD thesis, School of Informatics, University of Edinburgh, Scotland, 2007.

[12] J. Leidner, G. Sinclair, and B. Webber. Grounding Spatial Named Entities for Information Extraction and Question Answering. In *Proc. of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, 31–38, 2003.

[13] H. Li, R. K. Srihari, C. Niu, and W. Li. Location Normalization for Information Extraction. In *COLING'02*, 1–7, 2002.

[14] M. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling. STEWARD: Architecture of a Spatio-Textual Search Engine. In *GIS '07*, 186–193, 2007.

[15] T. Mandl, P. Carvalho, G. Di Nunzio, F. Gey, R. Larson, D. Santos, and C. Womser-Hacker. GeoCLEF 2008: The CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview. In *CLEF'08*, 808–821, 2008.

[16] I. Mani, J. Pustejovsky, and R. Gaizauskas, editors. *The Language of Time*. Oxford University Press, 2005.

[17] I. Mani and G. Wilson. Robust Temporal Processing of News. In *ACL '00*, 69–76, 2000.

[18] B. Martins, H. Manguinhas, and J. Borbinha. Extracting and Exploring the Geo-Temporal Semantics of Textual Resources. *Intl. Conf. on Semantic Computing*, 1–9, 2008.

[19] B. Martins, H. Manguinhas, J. Borbinha, and W. Siabato. A Geo-Temporal Information Extraction Service for Processing Descriptive Metadata in Digital Libraries. *e-Perimetron*, 4(1):25–37, 2009.

[20] MetaCarta Inc. *MetaCarta White Paper: MetaCarta GTS and MetaCarta GeoTagger.* `http://www.metacarta.com/resource-center-resources.htm`, 2008.

[21] A. Mikheev, M. Moens, and C. Grover. Named Entity Recognition without Gazetteers. In *EACL'09*, 1–8, 1999.

[22] D. Nadeau and S. Sekine. A Survey of Named Entity Recognition and Classification. *Linguisticae Investigationes*, 30(1):3–26, 2007.

[23] OpenCalais. `http://www.opencalais.com`.

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

[25] OpenStreetMap. `http://www.openstreetmap.org`.

[26] J. Pustejovsky, J. M. Castaño, R. Ingria, R. Sauri, R. Gaizauskas, A. Setzer, G. Katz, and D. Radev. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *IWCS-5*, 2003.

[27] J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo. The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics 2003*, 647–656, 2003.

[28] F. Schilder and C. Habel. From Temporal Expressions to Temporal Information: Semantic Tagging of News Messages. In *Proceedings of ACL'01 Workshop on Temporal and Spatial Information Processing*, 65–72, 2001.

[29] V. Tsotras. Recent Advances on Querying and Managing Trajectories. *Tutorial at the 10th Intl. Symposium on Spatial and Temporal Databases*, 2007.

[30] UIMA. `http://incubator.apache.org/uima/`.

[31] M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, G. Katz, and J. Pustejovsky. SemEval-2007 Task 15: TempEval Temporal Relation Identification. In *SemEval'07*, 75–80, 2007.

[32] M. Verhagen and J. Pustejovsky. Temporal Processing with the TARSQI Toolkit. In *COLING 2008: Companion Volume: Demonstrations*, 189–192, 2008.

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