A Temporal Text Mining Application in Competitive Intelligence

Haralampos Karanikas¹, George Koundourakis³, Ioannis Kopanakis², Thomas Mavroudakis¹

¹National & Kapodistrian Univ. of Athens, Knowledge Management Lab., Greece
{bkaranikas, tmavroudakis}@pha.uoa.gr,
²Technological Educational Institute of Crete, Heraklion, Greece
i.kopanakis@emark.teicrete.gr
³General Secretariat For Information Systems, Ministry Of Economy & Finance, Greece
g.koundour@gsis.gr

Abstract. In this paper we describe an application of our approach to temporal text mining in Competitive Intelligence for the biotechnology and pharmaceutical industry. The main objective is to identify changes and trends of associations among entities of interest that appear in text over time. Text Mining (TM) exploits information contained in textual data in various ways, including the type of analyses that are typically performed in Data Mining [17]. Information Extraction (IE) facilitates the semi-automatic creation of metadata repositories from text. Temporal Text mining combines Information Extraction and Data Mining techniques upon textual repositories and incorporates time and ontologies’ issues. It consists of three main phases; the Information Extraction phase, the ontology driven generalisation of templates and the discovery of associations over time. Treatment of the temporal dimension is essential to our approach since it influences both the annotation part (IE) of the system as well as the mining part.

1 Introduction and Motivation

We are living in the “Information Age” and biomedical knowledge is growing at an amazing speed, making comprehension a time-consuming and mentally heavy job for humans. The vast amount of data found in an organisation, some estimates run as high as 80%, are textual such as reports, emails, etc [12]. This type of unstructured data usually lacks metadata (data about data) and as a consequence there is no standard means to facilitate search, query and analysis. Knowledge discovery in text and Text Mining are mostly automated techniques that aim to discover high level information in huge amount of textual data and present it to the potential user (analyst, decision-maker, etc).

Text Mining elaborates unstructured textual information and examines it in an attempt to discover structure and implicit meanings “hidden” within the text [9] by utilising specialised data mining and Natural Language Processing (NLP) techniques operating on textual data. Text mining applications impose strong constraints on the usual NLP techniques [14] as the involvement of large volumes of textual data does not allow to integrate complex treatments.

Within our framework, Information Extraction is an essential phase in text processing. It facilitates the semi-automatic creation of (more or less) domain specific metadata repositories. Such repositories can be further processed using standard data mining techniques. The focus is to study the relationships and implications among entities (such as company, person etc.) and processes (such as business affairs etc)
participating in specific events. The goal is to discover important association rules over time within a document collection such that the presence of a set of entities or processes in a document implies the presence of another entity or process. For example, one might learn that whenever the company A occurs in news text, it is highly probable that a technology area is also mentioned for a specific time period.

The real world case study scenario demonstrated in this paper involves a company (www.biovista.com/) that specialises in corporate intelligence (including competitive intelligence - CI) products and services for the biotechnology and pharmaceutical industry. These products and services are targeted mainly to managers in charge of business development (and/or investment/strategic level issues) who wish to make decisions on industry and company developments or simply to monitor these on a regular basis.

In particular, our approach is applied to a number of documents that describe new alliances and partnerships of competitors. In an intelligent way, we discover business trends of competitors (or in the market in general). Competitors moving into markets, or the emergence of a new market, are of paramount importance. This type of analysis enhances knowledge acquisition regarding markets and competitors so that the business analyst would be able to support the proposed strategic plans of an/the organisation.

Our main contributions include:

- Integration on the database level of IE and DM in order to discover useful patterns from text.
- Incorporation of the time issue in the text discovery process.
- Incorporation of background knowledge via the use of ontologies in the discovery process.
- Metadata model for the text repository.
- An ontology driven technique for discovering temporal association in text.

The rest of the paper is organised as follows. In section 2 we discuss the process of analysing vast amount of textual data and provide the definitions that will be used throughout the paper. In the subsequent section 3 we demonstrate our approach, to perform temporal knowledge discovery in text, with the use of a realistic case study. Section 4 reviews related work in the field, while in section 5 we conclude and give some future research directions.

2 Knowledge Discovery in Text

Knowledge Discovery in Text (KDT) and Text Mining (TM) is an emerging research area that tries to resolve the problem of information overload by using techniques from data mining, machine learning, natural language processing (NLP), information retrieval (IR), information extraction (IE) and knowledge management. As with any emerging research area, there is still no established vocabulary for KDT and TM, a fact which can lead to confusion when attempting to compare results. Often, the two terms are used to denote the same thing. Knowledge discovery in text [6], [10] (or in textual databases), Text Data Mining [7] and Text Mining [14], [13], [16], [8], [15] are some of the terms that can be found in the literature.
We use the term KDT to indicate the overall process of turning unstructured textual data into high level information and knowledge, while the term Text Mining is used for the step of the KDT process that deals with the extraction of patterns from textual data. By extending the definition, for Knowledge Discovery in Databases (KDD) Fayyad and Piatetsky-Shapiro [5] we provide the following one:

*Knowledge Discovery in Text (KDT) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in unstructured textual data.*

KDT is a multi-step process, which includes all the tasks from the gathering of documents to the visualisation of the extracted information. A main goal of KDT is to make patterns understandable to humans in order to facilitate a better understanding of the underlying data [5].

*Text Mining (TM) is a step in the KDT process consisting of particular data mining and NLP algorithms that under some acceptable computational efficiency limitations produces a particular enumeration of patterns over a set of unstructured textual data.*

The above definition is better understood by comparing it with the related technology, Data mining. Data mining works with structured data, often numerical in nature. Text mining is analogous to data mining in that it uncovers relationships in information. Unlike data mining, it works with information stored in an unstructured collection of text documents.

The process of discovering knowledge in text includes three major stages [15], [11] (Fig. 1.):

1. **Collect relevant documents:** The first step is to identify what documents are to be retrieved. Once we have identified the source of our documents we need to retrieve the documents (from Web or form an internal file systems).
2. **Pre-processing documents:** This step includes any kind of transformation processes of the original documents retrieved. These transformations could aim at obtaining the desired representation of documents such as XML, SGML. The resulting documents are then processed to provide basic linguistic information about the content of each document. With an understanding of how words, phrases and sentences are structured, we can control text and extract information. From Text Mining perspective, our task in this pre-processing stage is to use language rules and word meanings to produce reusable by the other stages, representations of text. Main areas of linguistics include the Morphology (structure and form of words), Syntax (how words and phrases form sentences) and Semantics (the meaning of words and statements).
3. **Text mining operations:** High-level information is extracted (metadata creation). Patterns and relationships are discovered within the extracted information.
The accommodation of time into the KDT process provides the ability to discover richer knowledge which may also indicate cause and effect associations.

*Temporal Text Mining is an extension to Text Mining as it provides the capability of mining activity rather than just states upon textual data.*

The temporal issue influences mainly the NLP (Temporal annotations) and the mining (Temporal Data Mining) phases. NLP extracts and assigns the temporal features into textual objects and the mining techniques include the temporal dimensions introduced into the analysis.

### 3. Temporal Text Mining approach

The methodology we describe is a generic approach that can be applied to text databases of varying complexity. In our approach the process of detecting patterns within and across text documents (Text Mining) depends mainly upon information extraction techniques. By identifying key entities in a text along with the temporal characteristics, one can find relationships between terms and identify unknown links (‘hidden’) between related topics.

We consider that each fact (or event or template according to IE) defines a “transaction” that creates the associations between textual items. In analogy with a traditional data mining example, each fact is the corresponding market basket and the textual items (template slots according to IE) assigned on it constitute the items in the basket. Consequently our general goal of association extraction over time, given a set of facts, is to identify relationships between textual items and events such as the presence of one pattern implies the presence of another pattern over time.

The overall architecture of our system is shown in the following figure. First the IE module elaborates the input text in order to extract and fill the predefined template. Then the generalisation module performs the necessary generalisations upon specific slots of the template. Finally the filled templates feed the temporal association discovery module in order to extract the rules.
Metadata repository (and an associated model) is been utilised by our approach to store data that describe the information extracted from the documents. More specifically the role of the repository is to provide a neutral medium and unified representation forming the basis for analysis. Additionally it provides a more solid basis for large scale applications in text.

In the following sections we will demonstrate our methodology by presenting a sample of the case study, with some terms replaced due to confidentiality.

### 3.1 Preparation of the CI case study

For any modern organisation one of the most important issues nowadays, along with knowing its strengths/weaknesses and understanding its customers, is about knowing its competitors. Competitive intelligence is defined as “a systematic program for gathering and analysing information about your competitors’ activities and general business trends to further your own company’s goals” [19]. Due to the mainly unstructured format of the information sources, text mining is a valuable tool.

The real world case study scenario demonstrated in this paper involves a company (www.biovista.com/) that specialises in corporate intelligence (including competitive intelligence - CI) products and services for the biotechnology and pharmaceutical industry. Within Biovista, much of the information extraction process has historically been done manually. A human analyst has had to search a set of information sources gathering information, filtering, indexing and structuring it in an appropriate for use format. Examples of sources include press releases, scientific articles etc. The corpus of press releases for the case study has been derived from Business Wire (www.businesswire.com). Business Wire specialises in the electronic delivery of news releases and information directly from companies, institutions and agencies to the media, financial community and consumers. The press releases Biovista is interested in, concern business events within the life science domain.

Before we begin the process of gathering information about competitors we have to define some organising principles or else we can find ourselves searching, retrieving, and loading documents that do not meet our requirements.
The first task we have to accomplish in order to collect the relevant documents is to define the target competitors. In some industries i.e. automobile, this is a simple matter. In several other industries such as pharmaceutical is quite different. For example, while developing and testing new drugs is a complex process, which is often done by large companies, smaller companies are often formed to develop and market a single drug or use a single technology, such as genetic, in order to improve the manufacturing process for some class of drugs. This example shows that we should not only choose large competitor companies for a target company, but we should include and smaller companies. In fact, smaller companies in a specific sector may indicate a new dynamic market opportunities.

In our case study, few of the companies we have targeted include Combichem, Chromagen, Ono, Pharmacopeia, Pharmacia & Upjohn, Roche Bioscience etc. The documents we have gathered from business press releases on the internet concern the target companies and in particular discuss collaborations among them on specific scientific issues.

**Fig. 3. Sample collected document**

The 10th of January 1999 Pharmacopeia, Inc. (Nasdaq: PCOP) signed a new multi-year, multi-target research collaboration with Pharmacia & Upjohn Inc. Under the terms of the agreement, Pharmacopeia will screen its multi-million compound sample collection of diverse, small molecules against numerous targets chosen by Pharmacia & Upjohn each year. In addition to annual guaranteed payments, Pharmacia & Upjohn will make further payments based on the success of the screening programs. Optimization of active compounds identified by Pharmacopeia may be performed either by Pharmacia & Upjohn or by Pharmacopeia, for additional consideration. Financial terms of the agreement were not disclosed.

A *text mining goal* is an operational definition of the business goal. In other words, the text mining goal states project objectives in technical terms. For example, in our case study, one of the business goals might be: “Should we move in new technological areas?”. One of the sub-questions to answer is: “Are the competitors moving in new technological areas and which?”. The corresponding text mining goal might be: “According to the collaborations our competitors have signed the last year, discover relationships with new technological areas”.

### 3.2 Extracting information pieces

Following the problem definition and collecting the relevant documents is the phase of extracting information pieces from text. Information Extraction is the process of
identifying essential pieces of information within a text, mapping them to standard forms and extracting them for use in later processing [17]. In other words, it is the mapping of natural language texts (such as newswire reports, newspaper and journal articles, World Wide Web pages, etc.) into predefined, structured representation (or templates), which, when filled, represent an extract of key information from the original text [18]. The information refers to entities of interest in the application domain (e.g. companies or persons), or relations between such entities, usually in the form of facts (or events) in which the entities take part (e.g. company takeovers management successions). Once extracted, the information can then be stored, queried, mined, summarised in natural language, etc.

The IE component used in our framework is based on GATE system [3]. GATE allows a sequence of language processing components to work together and as a consequence marks up annotations on the input text. These language processing components include standard components such as sentence splitters, named entity recognisers etc. The result of the IE process is a set of named annotations over sections of the text.

Although the use of GATE was significant, especially for the named entity tasks, some of the work for the facts extraction been done manually due to the high precision required for the rest of the components. Another reason is that the focus of the current study, relating to IE is the outcome of this process.

In the current case study we consider extracting information pieces from a collection of documents that describe corporate activities (1998-1999). We are interested in extracting a specific fact that concerns Collaborations among companies on a specific technology subject and at some point in time. According to the sample document of figure 3, the output of the IE subsystem with the information pieces that we are interested on extracting, is shown in the following figure.

- **Company 1**: Pharmacopeia
- **Company 2**: Pharmacia & Upjohn
- **Event-Collaboration Type**: Research collaboration
- **Collaboration Subject**: Screen collection of molecules
- **Announced Date**: January 19, 1999
- **Valid Date**: January 10, 1999

![Fig. 4. Filled template with several slots that refer to collaboration facts.](image)

The above fact of interest that concerns collaborations among companies is defined by the following attributes: First of all, the two companies that participate. Then, the collaboration type and the collaboration subject that presents useful information of the relationship among the two companies. The collaboration subject usually includes technology terms of significance to the bio-knowledge worker. These technology terms point to technology areas/markets and they are utilised in order to represent the current or future directions of company activities. The last two features involve time. The announced date is the date that the fact was announced or published and the valid date is the actual date that the fact happened or is active. All these features have been defined in cooperation between the text mining analyst and the domain expert. Table 1 present case study examples of facts extracted.
Table 1. IE results of collaboration facts

<table>
<thead>
<tr>
<th>ID</th>
<th>Company1</th>
<th>Company2</th>
<th>Collab. Subject</th>
<th>DateAn</th>
<th>DateVal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pharmacopa</td>
<td>a</td>
<td>Pharmacia &amp; Upjohn</td>
<td>19/01/1999</td>
<td>10/01/1999</td>
</tr>
<tr>
<td>2</td>
<td>CombiChem</td>
<td>Chromagen</td>
<td>drug screening</td>
<td>15/03/1999</td>
<td>15/03/1999</td>
</tr>
<tr>
<td>3</td>
<td>CombiChem</td>
<td>Ono</td>
<td>drug discovery</td>
<td>04/01/1999</td>
<td>10/04/2000</td>
</tr>
<tr>
<td>4</td>
<td>CombiChem</td>
<td>Roche Bioscience</td>
<td>novel inhibitors</td>
<td>13/05/1998</td>
<td>05/06/1998</td>
</tr>
</tbody>
</table>

3.3 Metadata model

The information pieces extracted from documents are stored in a relational database according to our metadata model. The metadata model includes concepts derived from various existing standards such as the Dublin Core [23], Common Warehouse Metamodel [24], Message Understanding Conference (MUC7) [21], TimeML [22], Text Encoding Initiative TEI [20] etc.

Our metadata model is demonstrated using the linguistic notation and as such we will refer to annotations upon the documents. Where an annotation represents a form of meta-data attached to a particular section of document content. An annotation has a type (or a name) which is used to create classes of similar annotations, usually linked together by their semantics. Annotation set holds a number of annotations.

The most basic elements of information extraction metadata, in which we will refer as conceptual annotations, include Entities and Relations. Entities represent main points of a document and may be terms or concepts such as person names, locations, event’s action and other information pieces of interest within the documents. Relations describe the relationships between terms or concepts such as organizations, persons, places, etc. Combining entities and relationships we can define facts of interest within the document collection.

The set of annotations we are utilising is organised in three levels:

- **Structural Annotations.** This category of annotations is used to define the physical structure of the document (for example, the organisation into head, body, sections, paragraphs, sentences and tokens). This is particular useful when we have to guide our analysis in specific parts of the document. For example, if we are planning to utilise research papers in a knowledge discovery process, just working with the abstract sections could be enough.

- **Lexical Annotations.** These annotations are associated to a short span of text (smaller than a sentence), and identify lexical units that have some relevance for the framework: Named Entities, Terminology, Time Expressions, etc.

- **Semantic/Conceptual Annotations.** This type of annotations is not associated with any specific piece of text. They refer to lexical annotations. They (partially) correspond to what in MUC7 was termed 'Template Elements' and 'Template Relations' [21] (Template represents the final, tabular output format
of information extraction process). Ideally they should correspond to instances.

The temporal aspects of documents that are considered relevant to our analysis are captured with TimeML. Basically, TimeML uses five tags `<EVENT>`, `<TIMEX3>`, `<SIGNAL>`, `<LINK>` and `<DocCreationTime>`. The latter is self-explanatory, but in our model this information is captured by the Dublin core metadata Date which except Creation Date also includes Publication and Last modified Date. The scope of the annotation scheme comes directly from the interaction of the four remaining tags.

Two types of temporal entities are used: 1) Verbs, nominalizations, adjectives and prepositions that indicate something that happened are marked as `<EVENT>`. 2) Explicit references to times and calendar dates are marked as `<TIMEX3>`. How these two types of entities temporally interact is captured via `<SIGNAL>`, which marks explicit relations ("twice", "during"), and `<LINK>`, which marks implicit relations.

In the following figure the metadata model is presented.

![Fig. 5. Extracted information metadata, logical model](image-url)
3.4 Defining conceptual levels for analysis

An important role of ontologies in our approach is to be utilised by the user during the analysis (mining) step, in order to specify the conceptual levels of the discovered patterns. By utilizing generalisation rules based on business ontologies, Fig. 6 values of the template slots (extracted textual items i.e. CombiChem) are generalised at multiple conceptual levels.

![Fig. 6. Part of sample ontology](image)

As a result of this process, a set of slots is generated for every generalised slot of the original set. The generated attributes are instances of the same slots but in a different conceptual level (table 2). This facilitates the discovery of association patterns in various conceptual levels.

**Table 2. Generalised attributes**

<table>
<thead>
<tr>
<th>Template Slot</th>
<th>Generalised Template Slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Company</td>
<td>Bioscience sector that the company belongs</td>
</tr>
<tr>
<td>i.e. Pharmacopeia</td>
<td>Bioscience sector C i.e. Molecules sector</td>
</tr>
<tr>
<td>i.e. CombiChem</td>
<td>Bioscience sector A i.e. Drug sector</td>
</tr>
<tr>
<td>Collaboration subject</td>
<td>Generalised collaboration subject</td>
</tr>
<tr>
<td>i.e. Collection of molecules</td>
<td>i.e. Micro molecules sector</td>
</tr>
</tbody>
</table>

By applying the generalisations on table 1 that keeps the output of the IE we result with the tables 3 and 4. For attributes company and collaboration subject we get additional attributes with the generalised values.

**Table 3 Generalised collaborations (1/2)**

<table>
<thead>
<tr>
<th>ID</th>
<th>Company1</th>
<th>GenCompany1</th>
<th>Company2</th>
<th>GenCompany2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pharmacopeia</td>
<td>Bioscience Sector C</td>
<td>Pharmacia Upjohn</td>
<td>&amp; Bioscience Sector B</td>
</tr>
<tr>
<td>2</td>
<td>CombiChem</td>
<td>Bioscience Sector A</td>
<td>Chromagen</td>
<td>Bioscience Sector B</td>
</tr>
</tbody>
</table>
Table 4 Generalised collaborations (2/2)

<table>
<thead>
<tr>
<th>CollabSubject</th>
<th>GenCollabSubject</th>
<th>DateAn</th>
<th>DateVal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screen molecules</td>
<td>TechAreaB</td>
<td>19/01/1999</td>
<td>10/01/1999</td>
</tr>
<tr>
<td>drug screening</td>
<td>TechAreaA</td>
<td>15/03/1999</td>
<td>15/03/1999</td>
</tr>
<tr>
<td>drug discovery</td>
<td>TechAreaA</td>
<td>04/01/1999</td>
<td>10/04/2000</td>
</tr>
<tr>
<td>novel inhibitors</td>
<td>TechAreaB</td>
<td>13/05/1998</td>
<td>05/06/1998</td>
</tr>
</tbody>
</table>

Textual data set becomes more abstract and meaningful. The user is able to direct the analysis at the required conceptual levels and extract valuable knowledge that satisfy his needs.

Moreover, generalisation provides the necessary frequency for the patterns to be discovered. Usually, specific facts do not appear so frequently in the document collections that we keep. Generalisation of objects such as companies, products, etc. into market sector, technology area etc. respectively, provides frequent facts and thus the ability to detect useful patterns.

3.5 Discovering association over Time

Once we have extracted and stored in a database the information required from the documents we are able to perform temporal data mining algorithms in order to discover valuable temporal patterns.

The general goal of association extraction, given a set of data items, is to identify relationships between attributes and items such as the presence of one pattern implies the presence of another pattern [1]. Association rules upon textual items can be defined as follows:

Let $I = \{i_1, i_2, ..., i_m\}$ be a set of textual items (analogous to data item). A textual item varies from a single word to a multiword term i.e. person’s name, location, company’s name etc. Let $FDB$ (Facts Database) be a database of facts (analogous to transactions), where each fact $F$ consists of a set of textual items such that $F \subseteq I$. We use the term fact to refer to a set of textual items that refer to a specific piece of information that the user is interested. For example, the statement that “Sibilo acquired Oracle” is a fact and consists of the textual items: Sibilo (Company name), Acquire (Event) and Oracle (Company name). Given a textual itemset $X \subseteq I$, a fact $F$ contains $X$ only and only if $X \subseteq F$. An association rule is an implication of the form $X \Rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$ and $X \cap Y = \emptyset$. The association rule $X \Rightarrow Y$ holds in $FDB$ with confidence $c$ (supporting by the rule facts divided by the
facts that \( X \) is true) and has support \( s \) (supporting by the rule facts divided by the total facts). The task of mining association rules is to find all the association rules whose support is larger than a minimum support threshold and whose confidence is larger than a minimum confidence threshold.

An example of such associations may be the following: **The presence of Company A is associated with the presence of Technology Area B** (support \( s \), confidence \( c \)).

Each fact (collaborations) is treated individually in order to discover associations. However, in cases where fact histories exist, temporal patterns can be discovered. In our case study two temporal dimensions exist. The date that the fact announced and the actual date that the fact took place. For these specific temporal dimensions, the mining of the textual data set at different temporal periods results into a series of sets of association rules. Each set of association rules is the result of the mining process in one of the examined temporal periods. The strong association rules of one temporal period apply on the other temporal periods with the same, lower or higher support and confidence. It seems that the rule evolution in a temporal dimension is demonstrated by the fluctuation of its support and confidence in a series of temporal periods. After processing the result is of type: **In period \( t \) the presence of Company A and Company B is associated with the presence of Technology Area C** (support \( s \) and confidence \( c \) over time).

According to the text mining goal, as explained in section 3.1, specific queries expressed by potential knowledge workers are expressed upon the text repository (for examples tables 3, 4).

**Query:** “find all associations between a set of companies or the bioscience sector that the company belongs and any technology area for the period \( t \)”

**Result:** [Bayer AG] \( \rightarrow \) Combinational chemistry software [Rule stronger in March]

According to this application scenario the following tasks were performed:

1. User selects target attribute. Target attribute is the one for which there is interest on finding association rules. In our case we select the collaboration subject and the generalised collaboration subject, which represent the technology area that a specific biotechnology belongs.

2. User selects the attributes, generalised or not, that will appear in the left part of the rule. We select the attribute company 1 and the generalisation of the company 1 that represent the bioscience sector that the company belongs.

3. User must select a specific temporal dimension. In order to discover temporal rules, the algorithm utilises specific time slices that separates into continuous temporal periods of equal length. After the selection of the temporal attribute, the user should define the length of the temporal periods that are going to be examined. For example, this could be into minutes, hours, days etc. Finally the starting and ending points are defined. In our case we have selected the dimension of the date that the fact actually happened. Due to the nature of the news stories we had to deal, live news, also the date that the fact was announced could provide valuable information. Regarding the span of the temporal periods that we intend to examine, we have selected months.
We have chosen to present the most significant relations according to the measures of confidence and support. Additionally, we present relations that appear to be particularly important to the domain expert:

IF Company IS \textbf{CombiChem} THEN \textit{GenCollabSubject IS TechAreaA}

[Rule stronger at the beginning of the year and especially in March]

\begin{center}
\includegraphics[width=0.5\textwidth]{fig7}
\end{center}

\textbf{Fig. 7.} CombiChem $\rightarrow$ TechAreaA

IF GenCompany IS \textbf{Bioscience Sector A} THEN \textit{GenCollabSubject IS TechAreaA}

[Rule stronger at the end of the year and especially in November]

\begin{center}
\includegraphics[width=0.5\textwidth]{fig8}
\end{center}

\textbf{Fig. 8.} Bioscience Sector A $\rightarrow$ TechAreaA

IF GenCompany IS \textbf{Bioscience Sector B} THEN Collaboration Subject IS \textit{gene expression}

[Rule stronger at the mid and end of the year and especially in November]

\begin{center}
\includegraphics[width=0.5\textwidth]{fig9}
\end{center}

\textbf{Fig. 9.} Bioscience Sector B $\rightarrow$ Gene expression

IF Company IS \textbf{SmithKline Beecham} THEN Collaboration Subject IS \textit{high throughput screening}
The significance of the above type of results is obvious for the domain analyst. The analyst is able to detect movements of specific companies or bioscience sectors into technology areas over time. For a particular period of time the interest of companies for specific technologies can be discovered and as such, competitors.

The above approach is suitable for mining association rules and fluctuations of their strength, for textual items that appear in the same document. Additionally by using the method described in [31], we are able to discover association rules between textual items that appear in different documents. Via this technique, a data set of time series can be flattened by placing consecutive records in a single record. The number of records that are merged is called the time window size. Also the new attributes in the flattened record are renamed so as to avoid name clashes. By using this approach rules, we are able to mine association rules like the following:

\[
\text{IF at Time 1 GenCompany IS Bioscience Sector C and At Time 2 Company IS Inspire Pharmaceuticals, IDBS THEN at Time 5 GenCollabSubject IS TechAreaC}
\]

4. Related Work

There has been rather slow progress on research work in integrating IE techniques and data mining. Most existing systems tend to have a prescriptive structure: a number of topics are defined and the data is mined, or an atemporal one: a corpus of documents is mined for keywords, phrases, entities and relationships. [30] Most importantly, current techniques ignore most of the dynamic aspects, assuming that participating entities and formed links are static.

In 2000 we had discussed (to the best of our knowledge) the combination of information extraction and data mining [9], and the most relevant system to our approach is the one presented later the same year by Nahm and Mooney [13] which makes use of IE in order to construct a database of structured data from a document collection, and apply data mining upon them. However they are not involved with the temporal issue.

On the other hand research in text mining, taking into consideration temporal issues, is still very poor. [28] [29] At best altering mechanisms use simple temporal models that include either a sudden change in some rank, or the crossing of a threshold in order to determine the importance of a term over time. [30] TimeMines
(Construction Timelines with Statistical Models of Word Usage) automatically generates timelines by detecting documents that relate to a single topic (an event in a ‘history’). This is achieved exclusively by selecting and grouping semantic features in the text based on their statistical properties. [25] E-Analyst (Mining of Concurrent Text and Time Series) on the other hand, from University of Massachusetts at Amherst uses two types of data and attempts to find the relationship between them: the textual documents and numerical data, both with time stamps. The system discovers the trends in time series of numeric data (e.g., stock prices) and attempts to characterize the content of textual data (e.g. news articles) that precede these events. The objective is to use this information to predict the trends in the numeric data based on the content of textual documents that precede the trends. [26]

5. Conclusion and Future Work

The general goal of Text Mining is to automatically extract information from textual data. In this paper, we have presented an application of our approach for discovering associations in text over time. We have shown how the information extraction and the temporal features can be utilised in order to assist the mining process of associations over time. Additionally we have presented the use of generalisation for discovering association at a conceptual level the user desires and our metadata model for keeping the information required in the database.

We have made no attempt to provide a strong evaluation of the methods presented. This paper is describing the application of standard techniques to a textual data set in the context of a particular application – mining news for biotechnology and pharmaceutical industry. Evaluation, therefore, should be in terms of how this method improves some task. For example, does trends analysis of this sort improve the precision and the speed of alerting mechanisms?

We have used various components to demonstrate the usefulness of our approach. In the near future we are planning to apply our approach to large corpora. Another planned task is to extend the metadata model in order to capture the results of the mining phase for further analysis.

Acknowledgments

This research was supported by the Parmenides project “Ontology Driven Temporal Text Mining on Organisational Data for Extracting Temporal Valid Knowledge”, funded by the European Commission (ContractNo.IST-2001-39023, http://www.crim.co.umist.ac.uk/parmenides).

References

[8] Hehenberger M., Coupet P., “Text Mining applied to patent Analysis”.
[13] Nahm U.Y., Mooney R. J., “Using Information Extraction to Aid the Discovery of prediction Rules from Text”.
[16] Tan Ah-H., “Text Mining: The state of the art and the challenges”, in proceedings of the Pacific Asia Conf on Knowledge Discovery and Data Mining PAKDD’99 workshop on Knowledge Discovery from Advanced Databases.
[22] TimeML Annotation Guideline 1.00 (internal version 0.4.0), James Pustejovsky, Roser Sauri, Andrea Setzer, Rob Giauzas and Bob Ingria, July 21,2002 http://www.cs.brandeis.edu/%7Ejamesp/ar da/time/documentation/AnnotationGuide line-v0.4.0.pdf