Semantic Tagging of Domain-Specific Text Documents with DIAsDEM

Henner Graubitz*, Karsten Winkler*
Institute of Information Systems
School of Business and Economics
Humboldt-Universität zu Berlin
Spandauer Straße 1
D-10178 Berlin, Germany
{graubitz,kwinkler}@wiwi.hu-berlin.de

Myra Spiliopoulou
Institute of Business and
Technical Information Systems
Otto-v.-Guericke Universität Magdeburg
Universitätsplatz 2
D-39016 Magdeburg, Germany
myra@iti.cs.uni-magdeburg.de

Abstract

Large volumes of electronically available information are stored in textual form. The extraction of semantics from these documents and the characterization of their contents into a database-like schema is a necessary prerequisite for efficient search and for the fusion of documents semantically belonging together, be they documents about the same company, police reports or legal attests related to the same case.

In this study we present the approach pursued in the DIAsDEM[1] project to semantically tag documents and derive a corresponding XML document type definition. Our approach is based on iterative clustering of text units into homogeneous groups whose labels form the XML tags surrounding their contents and from which the document type definition for the collection is derived. DIAsDEM further incorporates a sophisticated mechanism for the preparation of texts on a specific knowledge domain that are characterized by jargon and syntax deviating from conventional linguistic rules. A case study on the semantic tagging of texts from the German Commercial Register archive demonstrates the advantages and applicability radius of our approach.

1 Introduction

For the past years, knowledge discovery in databases (KDD) has become a very active field both in research and practice. Aiming at the extraction and visualization of new, non-trivial, interesting and after all actionable knowledge from huge volumes of data, KDD combines various methods from statistics, machine learning, artificial intelligence and database research in a unifying process-centric framework [8]. Due to ubiquitous and easy-to-use Internet technologies as well as economic globalization, organizations frequently store mission-critical information in geographically distributed and often heterogeneous document collections and databases. The necessity of integrating all these data sources within a single, transparent and hence value-adding information system has constituted a significant challenge to the research area of information fusion [3].

In the research project DIAsDEM, we combine these two research directions by exploring KDD techniques for the integration of texts. Our main objective is the incorporation of legacy

*This work is partially funded by the German Research Society (DIAsDEM, DFG grant no. SP 572/4-1).

[1]Project funded by the German Research Society (DFG grant no. SP 572/4-1). DIAsDEM stands for "Datenintegration von Altlastdaten und semistrukturierten Dokumenten mit Mining Verfahren".

data and collections of semi-structured documents into an integrated information system that can be queried to support decision processes. There are three major steps to attain this objective: Firstly, semantic-carrying structure in semi-structured documents should be identified. Secondly, dependencies among the attributes of the legacy data of different sources must be detected. The results can afterwards be applied to integrate related data from various heterogeneous sources. In this study, we discuss the first issue.

KDD methodologies currently tend to focus on knowledge discovery in well structured, mostly relational data. However, a study mentioned by Tan indicated that approximately 80% of a company's information is contained in text documents [24]. Additionally, companies gradually realize that a purposeful management of both explicit and implicit organizational knowledge provides huge opportunities for creating sustainable competitive advantages. Obviously, text documents are one major source of explicit knowledge. Thus knowledge management requires techniques designed to extract knowledge from textual documents. In contrast to relational or object-oriented data frequently stored in databases, most text documents are not structured at all. Therefore successful knowledge discovery in texts requires a slightly different approach. Feldman and Dagan coined the phrase "knowledge discovery in textual databases" (KDT) that refers to the process of extracting useful knowledge from unstructured text documents [9].

In this paper, we present the DIAsDEM framework for semantic tagging of domain-specific text documents. Our goal is the enrichment of textual contents with metadata to efficiently facilitate search, browsing, querying, information integration and further in-depth knowledge discovery. Our approach enables both document exploitation for knowledge management and text preparation for information fusion by addressing the problem of making the semantics of a document explicit in the form of tags and deriving a document type definition that can play the role of a schema for the document collection. After successfully semi-structuring text documents in XML, advanced methodologies designed to handle semi-structured data can be applied to store, browse and query them in an efficient way. Transforming unstructured text into a semi-structured representation is also the first step to the integration of the source collection with other related data sources.

Related Work There are three categories of relevant work: Knowledge discovery in textual databases, research on semi-structured data and projects pursuing similar objectives. Tan briefly summarizes the current state of text mining and its future challenges [24]. Nahn and Mooney propose to combine methods from KDD and information extraction to perform text mining tasks [18]. Loh et al. suggest to exploit concepts rather than words for KDT purposes [14]. Mikheev and Finch describe a workbench for acquisition of domain knowledge from texts [17]. Feldman et al. propose text mining at the term level instead of focusing on words or linguistic tags [10]. In contrast, our approach uses text mining to discover semantics appropriate to serve as tags for text components. Our methodology enables knowledge discovery in domain-specific texts that significantly differ from average texts with respect to word frequency statistics.

Semi-structured data is another topic enjoying great interest within the database community [5, 1]. A lot of research effort has recently been put into methods inferring and representing existing structure in similar semi-structured documents [19, 25]. In order to transform existing content into XML documents, Sengupta and Purao propose a method that infers document type definitions using already tagged documents as input [23]. In contrast, we propose a method that tags the documents and derives a DTD for them. Most close to our approach is the work of Lumera, who uses keywords and rules to semi-automatically convert legacy data into XML documents [15]. However, his approach relies on the establishment of a rule base that drives the conversion, while DIAsDEM uses a KDD methodology in order to reduce human effort.

Bruder et al. introduce the search engine GETESS that supports query processing on texts by processing XML text abstracts. These abstracts contain language-independent, content-weighted summaries of domain-specific texts [4]. Decker et al. extract metadata from Web documents using the ontology-based system ONTOBROKER [6]. Embley et al. also apply ontologies to extract and structure information contained in data-rich unstructured documents [7]. Our approach is confined to the annotation of existing documents, retaining the original contents for further processing.

The rest of this paper is organized as follows: Section 2 gives an overview of the proposed framework for semantic tagging of domain-specific text collections. The following section 3 concisely describes all phases of the DIAsDEM process. Section 4 presents a case study that illustrates the application of the proposed framework to sematic tagging of textual records of one German Commercial Register. Finally, a summary and directions for future research are presented in section 5.

2 DIAsDEM Framework

In this paper, the notion of semantic tagging refers to the activity of annotating text documents with domain-specific XML tags. Rather than classifying entire documents or tagging single terms, the framework aims at annotating structural components of text documents that will be referred to as text units in the remaining sections. Reasonable text units may be sentences, paragraphs or even n-grams consisting of n subsequent words and sentences respectively. Table 1 illustrates this concept of semantic tagging: Each sentence is a text unit in this example.

<crime type="burglary" object="ring" company="Miller’s Jewelers Inc."> A platinum diamond ring was stolen from Miller’s Jewelers Inc. on Saturday in one of several thefts reported to police. </crime> <arrest person="Bryan Ray Owens"> The suspect Bryan Ray Owens was immediately arrested. </arrest> <value object="ring" money="3300 USD"> The ring was valued at $3,300 </object> <crime type="burglary" object="money order" money="1180 USD" date="28/03/2000"> In another incident, money orders worth $1,180 were stolen from a house in the 400 block of Bond Street on March 28, 2000. </crime>

Table 1: Example of semantically tagged sentences (police report)

KDT Process The discovery of semantic tags corresponds to obtaining useful knowledge about a collection of text documents. According to Mannila, knowledge discovery is an inherently interactive and iterative task. It should be seen as a process supported by an interactive KDD system [16]. DIAsDEM adopts this process-centric approach and therefore contains a complex process for discovering knowledge in textual databases. The proposed KDT process is depicted in Figure 1. It illustrates the main phases of iterative knowledge discovery and outlines necessary activities.

We propose a methodology for semi-automatically converting natural language text documents into semantically annotated XML documents. To attain our objectives, tags describing semantically similar text units (e.g., sentences) must be found, an XML document type definition representing a coarse schema of the source collection must afterwards be generated and text units finally have to be annotated using the discovered tag set. Additionally, named entities (e.g., names or dates) could be extracted from text units in order to serve as attributes of XML tags that highly increase the information value of the corresponding metadata. The reduction of human effort through support for the knowledge engineer is also of great importance.

Areas of Application In DIAsDEM we concentrate on the semantic tagging of similar text documents originating from a common domain. Firstly, the proposed KDT process utilizes clustering techniques to find groups of text units that have both a high intra-cluster and a low inter-cluster similarity. However, the exploratory analysis of data using clustering algorithms is only suitable if there is at least a cluster tendency within the data domain [12]. It is expected that

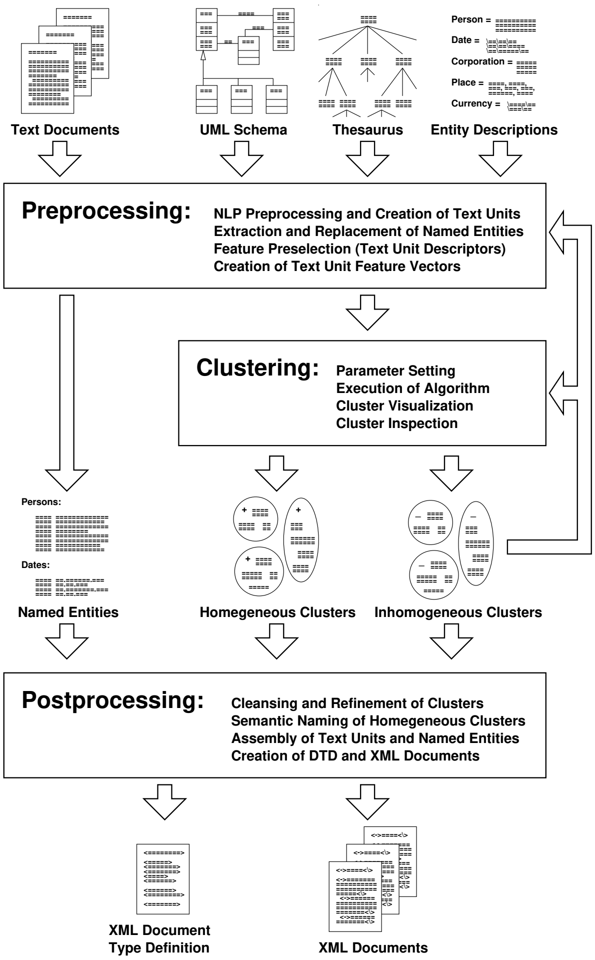

Figure 1: KDT process of the DIAsDEM framework for semantic tagging

text units of domain-specific documents are more likely to meet this criterion than text units originating from diverse application domains. Secondly and analogously to [20], the creation of expressive XML tags requires pure as well as complete clusters of text units to be appropriately named. Pure, i.e. homogeneous, clusters do not contain inappropriate text units, whereas complete clusters contain all text units that accord with the respective topic. To meet this requirement, the preprocessing phase among other things deploys a controlled vocabulary for dimension reduction that typically contains application-specific jargon. The imposed restriction on the documents' domain largely facilitates dimension reduction as well as the maintenance of the corresponding controlled vocabulary. Finally, creating and after all using an XML document type definition should only be worthwhile for a certain type of text documents from a common domain.

Semantic tagging of unstructured, natural language texts often significantly increases the value of text collections and enables the deployment of value-adding services superior to conventional full-text search. Despite the focus on domain-specific text documents, there are many possible applications of the proposed framework covering different fields of information technology. The DIAsDEM approach is thus appropriate for semantically tagging archives of public announcements of courts and administrative authorities, quarterly and annual reports to shareholders of public companies, textual patient records in health care applications, open question interview data collected in opinion polls as well as product and service descriptions published on electronic marketplaces.

We reduce the human effort by performing iterative clustering: Only the homogeneous clusters finally retained must be inspected. Also, cluster homogeneity ensures that the selection of appropriate cluster labels will be low. However, the amount of manual effort tends to be rather independent of the exact number of individual documents contained in the source collection. It is proportional to the number of discovered clusters and not to the number of processed text units. Therefore it is assumed that the benefits of semantic tagging more easily outweigh the corresponding costs in large collections of similar text documents. Particularly, organizations could effectively refine huge amounts of accumulated textual legacy data by adding metadata through semantic XML tagging.

3 Iterative KDT Process

After setting the objectives and prerequisites of the DIAsDEM framework for semantic tagging, this section concisely introduces the corresponding KDT process. Constituting the main input data, all source text documents must be composed of either plain or structurally annotated text. In the latter case, a clearly defined markup language (e.g., SGML) should be used to mark the structural components of each text document such as sections, paragraphs and sentences. Due to the widespread usage of multimedia documents that are composed of texts, images, charts or even sound and video sequences, relevant texts must often be extracted from proprietary file formats using individually tailored wrappers or semi-automatic extraction tools such as NoDoSE [2].

The KDT process incorporates domain knowledge supplied by knowledge engineers in order to ensure a high quality of tagging. Firstly, the basis for successful feature selection is a controlled vocabulary comprising of specific nomenclature as well as other terminology with sufficiently high discriminatory power. Although requiring large efforts for creation and maintenance, thesauri and ontologies seem to be suitable methodologies for that purpose. Domain knowledge is also reflected in descriptions of named entities to be extracted from text documents. Domain-specific algorithms, patterns or regular expressions are necessary to successfully extract for example names of people, companies or amounts of money from text documents. The conceptual model of the specific domain is an optional but highly recommended input that also incorporates essential knowledge. The Unified Modeling Language (UML) provides various object-oriented diagrams that help to gain a deep insight into the domain described by the source document collection. UML class diagrams depict domain-specific classes and their relationships. They contain real-world names of classes, attributes, operations and relationships and might therefore be used to reveal potentially important terminology.

Preprocessing In accordance with Mannila's KDD process, preparing the data set is the second phase in the process of discovering useful knowledge. Knowledge discovery in textual databases requires a particularly extensive data preprocessing phase in order to reduce the huge dimensionality of natural language texts caused by the immense syntactic, semantic and pragmatic variety of language written or spoken by humans. Even though computing power is increasing, characterizing a text by a set of attributes consisting of all occurring words is still computationally prohibitive. Since our goal is to identify clusters of semantically similar text units in order to assign them appropriate XML tags, preprocessing starts with defining the level of text unit granularity. This choice must be considered very carefully because only these text units will later be semantically annotated. In our approach, XML tags contain a semantic label as well as named entities extracted from the corresponding text unit.

Data preprocessing includes a linguistic tokenization phase that also separates text units from each other as well as a stemming phase that reduces all words to their canonical forms. The source document collection is afterwards transformed into a collection of text units. Additionally, named entities can be extracted from each text unit. They are replaced by individual placeholders. In this framework, each preprocessed text unit is thereafter converted into a so-called text unit vector. The vector space model is a feature representation for text documents originally developed within the SMART information retrieval project [21]. Each text unit vector contains all attribute values of the corresponding text unit. They can be created using a variety of methods that only depend on the choice of attributes describing a text unit. We principally propose the use of attributes reflecting the existence of domain-specific terms (so-called descriptors) contained in the controlled vocabulary. In contrast to selecting all canonical word forms to be attributes, our approach drastically reduces the vector dimension and thus facilitates the clustering phase.

Iterative Clustering Pattern discovery is the objective of the third phase of the generic KDD process suggested by Mannila. In DIAsDEM, pattern discovery aims at partially automated cluster labeling and DTD derivation for the document collection. At this stage, an explorative clustering technique is thus used to group semantically similar text units. Cluster analysis is often described as the art of finding groups in data [13]. Clustering can informally be defined as the unsupervised classification of patterns into groups based on similarity [12]. This task has been successfully addressed within different research communities such as statistics, information retrieval and computer science. For that reason, there are various algorithms publicly available that reflect distinguished concepts and methodologies of cluster analysis as well as varying fields of applications.

At this point in time, we thus refrain from adding yet another clustering algorithm to the large toolbox. We rather suggest iterative execution of the selected clustering algorithm: In each iteration loop, all discovered clusters are visualized and evaluated to support the knowledge engineer in the activity of identifying homogeneous text unit clusters. The collection of text unit vectors to be clustered in the next loop only consists of vectors that are contained in inhomogeneous clusters containing no semantically similar text units at all. Furthermore, the remaining text unit vectors may be altered to represent slightly different features and the parameters of the clustering algorithm may also be adjusted. The knowledge engineer will have to decide when to stop the iteration. Our experiments revealed that this iterative process largely improves the quality of the tagging process.

Postprocessing The phase prior to the last in both Mannila's generic KDD process and the DIAsDEM framework is devoted to postprocessing of discovered patterns. As a result of the previous clustering phase, two different types of text unit clusters must be distinguished. There are clusters containing mostly homogeneous text units on the one hand. After cleaning and refining the contents of these clusters, they are assigned a semantic name by the knowledge engineer. Each cluster's name should reflect concepts and topics being present in the corresponding text units. This task must be strongly supported by the system in order to minimize human effort. On the other hand, the remaining clusters only contain rather inhomogeneous text units. Along with text units not assigned to any cluster at all, the contents of those clusters cannot be semantically annotated due to the obvious lack of common topics.

In order to attain the objective of semantic tagging, every text unit contained in a semantically named cluster is annotated with an XML tag being identical to its cluster's name. Depending on the knowledge engineer's decision, XML tags may also contain attributes that will be equivalent to previously extracted named entities of the corresponding text unit. According to their semantics within the concepts described by an XML tag, attributes may be assigned a semantic name as well. All named entities denoting persons, companies and amounts of money are stored in a single attribute by default.

XML Tagging Finally, XML documents are created by replacing the extensively preprocessed contents of all XML tags (i.e. tokenized and stemmed text units) as well as all untagged text units by their natural language counterparts. Each XML document must exactly match its corresponding text document with regard to contents and order of text units. Therefore both the collections of annotated and untagged text units must afterwards be re-arranged in order to create valid XML documents that strictly correspond to their counterparts in the source collection. Additionally, a document type definition for all XML documents contained in the resulting collection is derived. Document type definitions generated by the current prototype only consist of valid XML tags, their attributes and placeholders for textual contents. Information about the semantic structure of the document collection must later be inferred using appropriate algorithms.

As a result of applying the proposed framework, the source collection of domain-specific text documents is converted into the corresponding collection of semantically annotated XML documents. The specific users' needs determine the deployment of further algorithms, techniques and applications that either utilize the existence of semantic metadata within XML documents or exploit the information contained in the document type definition.

4 Case Study

We have applied the iterative KDT process of DIAsDEM to semantically tag a collection of text documents originating from the German Commercial Register of the Potsdam district court. Such a register contains important information about the legal affairs of the companies existing in the court's district. Companies are obliged by the German Commercial Code to submit information about their affairs such as the establishment of new branch offices, changes of their share capital and managerial head as well as mergers and acquisitions. The importance of this register is stressed by the fact that its entries have both a right-confirmation and right-generating effect according to law.

HRB 12375 - 16.03.1999	GEBATEL Gesellschaft für Bahntelekommunikation mbH (Edisonstraße 6, 14612 Falkensee).	publiziert am 19.03.1999

Die Planung, Projektierung und der Vertrieb von auch für den Einsatz in Bahnen geeigneten Telekommunikationsanlagen. Stammkapital: 25.000 EUR. Gesellschaft mit beschränkter Haftung. Der Gesellschaftsvertrag ist am 22. Februar 1999 abgeschlossen. Ist nur ein Geschäftsführer bestellt, so vertritt er die Gesellschaft allein. Sind mehrere Geschäftsführer bestellt, so wird die Gesellschaft durch zwei Geschäftsführer gemeinschaftlich oder durch einen Geschäftsführer in Gemeinschaft mit einem Prokuristen vertreten. Einzelvertretungsbefugnis kann erteilt werden. Horst Peter Dürbeck, geb. am 27.06.1943, Berlin, und Thomas Hach, geb. am 16.05.1957, Falkensee, sind zu Geschäftsführern bestellt. Sie vertreten die Gesellschaft jeweils einzeln und sind befugt, Rechtsgeschäfte mit sich selbst oder mit sich als Vertreter Dritter abzuschließen. Nicht eingetragen: Die Bekanntmachungen der Gesellschaft erfolgen im Bundesanzeiger.

Table 2: Example entry in a German Commercial Register

To attain this purpose, most information contained in Commercial Register records is made available to the general public. It is published by both nationwide and local newspapers. Additionally, anyone can inspect a court's Commercial Register without having to prove a certain interest. As table 2 illustrates, each entry consists of a structured part and an unstructured textual section. The former contains relational data such as the company's registered name, its record number as an identifier, the business address as well as relevant dates of registration and publication. This information can easily be extracted using wrapper technologies. The unstructured part of each entry contains the registered text as recorded by the court's clerks. In most cases, this text consists of up to ten sentences describing the fact to be published. There are three major categories of entries: Foundation entries for new companies, update entries and entries announcing that a company closes.

Preprocessing The framework was applied to a collection of 1.145 Commercial Register entries published by the district court Potsdam[2]. This collection includes all entries related to foundations of new companies in 1999 and was selected because it demonstrates the ideal domain of DIAsDEM: Firstly, it is a rather large collection of domain-specific similar text documents. Secondly, our intuition had revealed that sentences exhibited a certain cluster tendency. Thus representing each sentence as a text unit was justified. As a result of this decision regarding the level of granularity, each XML tag always annotates a single sentence. Finally, the documents are written in a special jargon not conforming to syntactic and semantic rules of NLP and thus demonstrating the need for a specialized vocabulary.

The conceptual model reflecting the domain includes two UML class diagrams depicted in Figures 2 and 3 respectively. They reveal important concepts as well as special business and judicial terminology. The modeling phase also created the basis for defining a controlled vocabulary. This vocabulary was subsequently extended to develop a domain-specific thesaurus containing a hierarchy of 70 relevant descriptors and 109 non-descriptors. Currently, thesaurus generation is supported by word frequency statistics and a thesaurus editor.

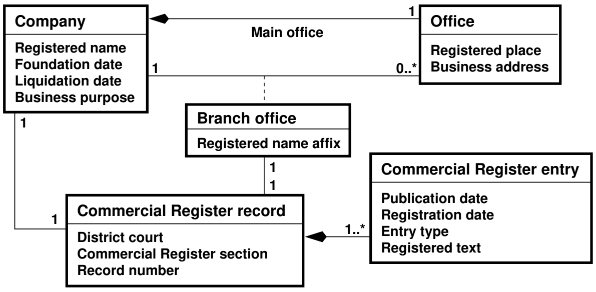

Figure 2: UML class diagram describing the domain context

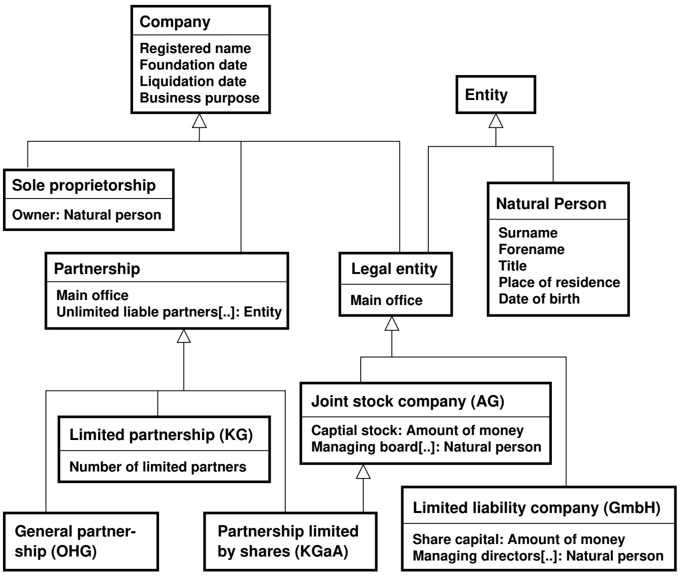

Figure 3: UML class diagram describing the taxonomy of German company types

[2]`http://www.amtsgericht-potsdam.org`

After extracting the relevant texts from HTML source files, all text documents have been tokenized to separate text units from each other. Afterwards, a part–of–speech tagger (TreeTagger) capable of processing German texts was deployed to create lemma forms and thereby reducing the number of unique word forms from 10.613 to approx. 5.400 [22]. Due to the absence of German extractors of named entities, a small but powerful pattern-based Java application has been developed to extract various, previously defined entities from text units. The Java-based DIAsDEM workbench was used to determine approx. 85 text unit descriptors and to finally create 10.785 text units vectors.

Clustering In the context of this case study, explorative pattern discovery by means of clustering had to be applied to detect groups of semantically similar sentences. At this stage, the IBM Intelligent Miner for Data was employed[3]. Its demographic clustering mining function first determines similarities between text unit vectors and afterwards defines clusters that maximize the value of Condorcet's criterion. This criterion is the difference of the sum of pair-similarities for all text unit vectors in the same cluster and the sum of all pair-similarities for text unit vectors in different clusters. The number of clusters to be generated is automatically determined by the miner [11].

According to the framework, the collection of text unit vectors was iteratively clustered. After each of six iterations, the contents of all created clusters were visualized by temporarily transforming vectors into the corresponding natural language sentences. Supported by descriptive cluster statistics (e.g., relative frequency of descriptor occurrence), a domain specialist evaluated the clusters with respect to semantic similarity of their contents. All homogeneous clusters were put aside for labeling. The remaining inhomogeneous clusters constituted the input to the next clustering loop. Although possible, we refrained from modifying the set of text unit descriptors during the iterative clustering phase. After six iterations of clustering, 91 homogeneous clusters were successfully detected representing approx. 96% of all text units to be clustered.

Postprocessing In the last phase, all sentences of the document collection are tagged with labels of the clusters they belong to. Usually, the semantic name of a cluster includes a combination of frequently occurring text unit descriptors. This fact was exploited by the DIAsDEM workbench to automatically generate default XML tags for every cluster. The approach drastically reduced necessary human efforts.

In order to transform source texts into the corresponding collection of XML documents, each text unit contained in a homogeneous cluster was tagged with the cluster's label. Subsequently, both tagged and untagged text units contained in inhomogeneous clusters were merged to create XML documents being equivalent to their text counterparts. XML tags of sentences that contained extracted named entities (e.g., persons) were extended to include their values as additional attributes. Table 3 exemplary illustrates the XML document that was eventually created after processing the textual section shown in Table 2. Finally a simple document type definition (DTD) was automatically derived. It describes the semantic structure of the resulting XML collection coarsely. Table 4 contains an excerpt.

Due to the intense business demand for this commercial information, there are several information brokers offering both online and offline services to retrieve relevant knowledge from Commercial Registers. However, the current state of service only encompasses SQL queries to access relational data and full-text queries to search in textual sections. The proposed DIAsDEM framework now provides a powerful method to enhance the quality and therefore increase the value of Commercial Register data. After applying the proposed framework to the relevant document collections, users will be able to search annotated documents by utilizing XML tags and their attributes in queries. Tagged documents may also serve as preprocessed input to further KDT efforts aimed at obtaining useful knowledge about companies. The discovered XML document type definition of Commercial Register entries can be used to facilitate information fusion.

[3] `http://www.ibm.com/software/data/iminer`

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE CommercialRegisterEntry SYSTEM 'CommercialRegisterEntry.dtd'>

<CommercialRegisterEntry> <BusinessPurpose> Die Planung, Projektierung und der
Vertrieb von auch für den Einsatz in Bahnen geeigneten Telekommunikationsanlagen.
</BusinessPurpose> <ShareCapital NE="AmoutOfMoney=[25000 EUR]"> Stammkapital:
25.000 EUR. </ShareCapital> <LimitedLiabilityCompany> Gesellschaft mit be- schränkter
Haftung. </LimitedLiabilityCompany> <ConclusionOfPartnershipAgreement
NE="Date=[22.02.1999]"> Der Gesellschaftsvertrag ist am 22. Februar 1999 abgeschlossen.
</ConclusionOfPartnershipAgreement> (...) <ManagingDirectorAppointment
NE="Person=[Dürbeck; Peter; 27.06.1943; Berlin], Person=[Hach; Thomas;
16.05.1957; Falkensee]"> Horst Peter Dürbeck, geb. am 27.06.1943, Berlin, und Thomas Hach, geb.
am 16.05.1957, Falkensee, sind zu Geschäftsführern bestellt. </ManagingDirectorAppointment>
<ExtentOfRepresenationRight> Sie vertreten die Gesellschaft jeweils einzeln und sind befugt,
Rechtsgeschäfte mit sich selbst oder mit sich als Vertreter Dritter abzuschließen.
</ExtentOfRepresenationRight> <PublicationsOfCompany> Nicht eingetragen: Die
Bekanntmachungen der Gesellschaft erfolgen im Bundesanzeiger. </PublicationsOfCompany>
</CommercialRegisterEntry>
```

Table 3: XML document representing an annotated Commercial Register entry

```
<?xml version="1.0" encoding="ISO-8859-1"?>

<!ELEMENT CommercialRegisterEntry ( #PCDATA | FoundationCompany |
ExtentOfRepresentationRight | ShareCapital | NumberOfLimitedPartners |
MainOfficeTransfer | (...) | LimitedLiabilityCompany )* >

<!ELEMENT FoundationCompany (#PCDATA)>
<!ELEMENT ExtentOfRepresentationRight (#PCDATA)> (...)
<!ELEMENT LimitedLiabilityCompany (#PCDATA)>

<!ATTLIST FoundationCompany NE CDATA #IMPLIED>
<!ATTLIST ExtentOfRepresentationRight NE CDATA #IMPLIED> (...)
<!ATTLIST LimitedLiabilityCompany NE CDATA #IMPLIED>
```

Table 4: XML document type definition of Commercial Register entries

5 Conclusion

Collections of unstructured text documents as well as textual legacy data often contain information of great potential value. Nevertheless, text documents are frequently either stored in hardly accessible archives or they can only be queried by unsatisfactory full-text search. Knowledge discovery in texts is an enabling methodology to transform textual contents into valuable assets.

In this paper we have introduced the DIAsDEM framework for semantic annotation of domain-specific texts with an iterative KDT process. We used conventional clustering methods in a new context. The iterative clustering mechanism groups similar sentences together, derives cluster labels in a partially automated subprocess, tags documents with these labels and derives a DTD describing the document collection. Instead of focusing on document clustering or word tagging, various structural parts of texts can be entities to be semantically annotated. Additionally, a named entity extractor identifies and tags person names, dates, amounts of money etc. within each sentence. In this study, we have concentrated on the former, i.e. the iterative clustering mechanism. To attain this goal, concepts and methodologies from KDD, natural language processing, information retrieval and information extraction were brought together in a unifying framework. Our approach incorporates domain knowledge by utilizing conceptual models and thesauri.

The proposed framework has been successfully applied to a collection of publicly available

Commercial Register entries. A Java-based DIAsDEM workbench has been implemented to support the entire KDT process. Of course, many open issues remain: First of all, the DTD extraction method from cluster labels should be refined to reflect the complexity of the documents. The results of schema discovery should afterwards be used to design enhanced query services that are for instance capable of identifying documents on the same subject in different data sources. The discovered schema of Commercial Register entries will be used to facilitate information integration with other related data sources.

Additionally, advanced methods of natural language processing can be deployed in the preprocessing phase to automate thesaurus generation and to identify grammatical structures that might serve as text units (e.g., noun phrases and subordinate clauses). The human efforts necessary to select a promising set of text unit descriptors must be reduced by applying appropriate KDD techniques. Therefore a semi-automatic feature selection method will be developed that suggests a minimal set of highly discriminating descriptors. The clustering phase offers challenges as well: Existing clustering algorithms must be evaluated with respect to the objectives of this framework. To measure the homogeneity of clusters, an appropriate evaluation metric will also be developed.

References

[1] S. Abiteboul, P. Buneman, and D. Suciu. *Data on the Web: From Relations to Semistructured Data and XML.* Morgan Kaufman Publishers, San Francisco, 2000.

[2] B. Adelberg. NoDoSE - a tool for semi-automatically extracting semi-structured data from text document. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 283–294, Seattle, WA, USA, June 1998.

[3] H. Arabnia and D. D. Zhu, editors. *Proceedings of the First International Conference on Multisource-Multisensor Information Fusion*, Las Vegas, NV, USA, July 1998. CSREA Press.

[4] M. Becker, J. Bedersdorfer, I. Bruder, A. Düsterhöft, and G. Neumann. GETESS: Constructing a linguistic search index for an Internet search engine. In *Proceedings of the 5th International Conference on Applications of Natural Language to Information Systems*, Versailles, France, June 2000.

[5] P. Buneman. Semistructured data. In *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 117–121, Tucson, AZ, USA, May 1997.

[6] S. Decker, M. Erdmann, D. Fensel, and R. Studer. ONTOBROKER: Ontology based access to distributed and semi-structured information. In R. e. a. Meersman, editor, *Database Semantics: Semantic Issues in Multimedia Systems, Proceedings TC2/WG 2.6 8th Working Conference on Database Semantics*, Rotorua, New Zealand, 1999.

[7] D. W. Embley, D. M. Campbell, R. D. Smith, and S. W. Liddle. Ontology-based extraction and structuring of information from data-rich unstructured documents. In *Proceedings of the 1998 ACM 7th International Conference on Information and Knowledge Management*, pages 52–59, Bethesda, MD, USA, 1998.

[8] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34, November 1996.

[9] R. Feldman and I. Dagan. Knowledge discovery in textual databases (KDT). In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, pages 112–117, Montreal, Canada, August 1995. AAAI Press.

[10] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler, and O. Zamir. Text mining at the term level. In *Proceedings of the Second European Symposium*

on Principles of Data Mining and Knowledge Discovery, pages 65–73, Nantes, France, September 1998.

[11] IBM DB2 Intelligent Miner for Data: Using the Intelligent Miner for Data. IBM Corporation, 1999.

[12] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, September 1999.

[13] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York, Chichester, Brisbane, Toronto, Singapore, 1990.

[14] S. Loh, L. K. Wives, and J. P. M. d. Oliveira. Concept-based knowledge discovery in texts extracted from the Web. *ACM SIGKDD Explorations*, 2(1):29–39, 2000.

[15] J. Lumera. Große Mengen an Altdaten stehen XML-Umstieg im Weg. *Computerwoche*, 27(16):52–53, 2000.

[16] H. Mannila. Methods and problems in data mining. In *Proceedings of the International Conference on Database Theory*, pages 41–55, Delphi, Greece, January 1997.

[17] A. Mikheev and S. Finch. A workbench for acquisition of ontological knowledge from natural language. In *Proceedings of the Seventh conference of the European Chapter for Computational Linguistics*, pages 194–201, Dublin, Ireland, March 1995.

[18] U. Y. Nahm and R. J. Mooney. Using information extraction to aid the discovery of prediction rules from text. In *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD-2000) Workshop on Text Mining*, pages 51–58, Boston, MA, USA, August 2000.

[19] S. Nestrov, S. Abiteboul, and R. Motwani. Inferring structure in semi-structured data. *SIGMOD Record*, 26(4):39–43, 1997.

[20] M. Perkowitz and O. Etzioni. Adaptive Web sites. *Communications of the ACM*, 43(8):152–158, August 2000.

[21] G. Salton, J. Allan, and C. Buckley. Automatic structuring and retrieval of large text files. *Communications of the ACM*, 37(2):97–108, February 1994.

[22] H. Schmid. Probabilistic part–of–speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK, September 1994.

[23] A. Sengupta and S. Purao. Transitioning existing content: Inferring organization-spezific document structures. In K. Turowski and K. J. Fellner, editors, *Tagungsband der 1. Deutschen Tagung XML 2000, XML Meets Business*, pages 130–135, Heidelberg, Germany, May 2000.

[24] A.-H. Tan. Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, pages 65–70, Beijing, China, April 1999.

[25] K. Wang and H. Liu. Discovering structural association of semistructured data. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):353–371, May/June 2000.