SEMI-AUTOMATED XML TAGGING OF PUBLIC TEXT ARCHIVES: A CASE STUDY

Karsten Winkler and Myra Spiliopoulou

Leipzig Graduate School of Management, Leipzig, Germany
e-mail: {kwinkler,myra}@ebusiness.hhl.de

Abstract

Public archives contain large and continuously growing volumes of electronically available text documents. In many countries, public authorities are required by law to publish certain data to satisfy the information needs of the general public. In contrast to plain text documents, semantically tagged XML documents along with appropriate query languages largely facilitate searching and browsing in public archives for interested citizens. However, transforming textual legacy data into semantically annotated XML documents should be automated to minimize costly human effort. In this paper, we present the DIAsDEM framework for semi-automated semantic tagging of domain-specific text documents in a case study. Our framework includes a complex knowledge discovery process that groups structural text units (e.g., sentences) based on similarity of their content, derives semantic labels for qualitatively acceptable clusters, semantically tags text units and derives a preliminary unstructured XML DTD for the archive. We apply this framework to collections of publicly available Commercial Register archives and finally evaluate the quality of our approach.

Keywords: XML, semantic tagging, DTD derivation, data mining, text mining, clustering

INTRODUCTION

Public authorities have been accumulating huge amounts of text archives due to legal obligations for the past decades. Many textual and thus unstructured archives such as bulletins of administrative authorities, environmental reports and announcements of courts must be published according to law. In some countries, authorities slowly turn towards e-driven service providers to meet the intense information demands of their citizens. Unfortunately, most data currently resides in unstructured text archives, lacks meta-data and is only accessible through limited search mechanisms (i.e. full-text search).

The successfully emerged Extensible Markup Language XML combined with appropriate query languages is capable of solving these challenges by providing a solid basis for efficient and Web-based information systems in the public administration sector. In contrast to plain texts or HTML documents, semantically annotated XML documents do not only facilitate efficient searching, browsing, querying and information integration. The markup language XML is accompanied by an elaborated set of co-standards for visualization, i.e. the Extensible Stylesheet Language XSL. Transforming legacy archives into XML collections

hence enables information providers to simultaneously publish the same data via different media channels, such as HTML, PDF or special formats for handicapped persons.

Unfortunately, users tend to dislike creating semantic meta-data due to the necessary effort involved. In this paper, we summarize the DIAsDEM framework for semantic tagging of domain-specific text documents and report on the application of this framework to an archive of publicly available text documents. Our objective is the semi-automated annotation of textual documents to minimize costly human effort. Therefore, we utilize a knowledge discovery in databases (KDD) methodology. Aiming at the extraction of new, non-trivial, interesting and after all actionable knowledge from huge volumes of data in a process-centric framework [FaPiSm1996], KDD is an active field in research and practice. In this study, we focus on the application of the DIAsDEM framework in a case study.

The rest of this paper is organized as follows: Related work is discussed in the next section. The third section summarizes our proposed DIAsDEM framework for semantic tagging of domain-specific texts. Thereafter, we present a case study in which this framework has been successfully applied to the public German Commercial Register. Finally, we conclude and look at future research and development issues.

RELATED WORK

Nahn and Mooney propose the combination of methods from KDD and information extraction to perform text mining tasks [NaMo2000]. They apply standard KDD techniques to a collection of structured records that contain previously extracted, application-specific features from texts. Feldman et al. propose text mining at the term level instead of focusing on linguistically tagged words [FFKL+1998]. The authors represent each document by a set of terms and additionally construct a taxonomy of terms. The resulting dataset is input to KDD algorithms such as association rule discovery. Our DIAsDEM framework adopts the idea of representing texts by terms and concepts. However, our goal is the semantic tagging of structural text units (e.g., sentences or paragraphs) within the document according to a global DTD and not the characterization of the entire document's content. Loh et al. suggest to extract concepts rather than individual words for subsequent use in KDD efforts at the document level [LoWiOl2000]. Similarly to our framework, the authors suggest to exploit existing vocabularies such as thesauri for concept extraction. Mikheev and Finch describe a workbench to acquire domain knowledge from texts [MiFi1995]. As the DIAsDEM Workbench, their approach combines methods from different fields of research in a unifying framework.

Our approach shares with this research thread the objective of extracting semantic concepts from texts. However, concepts to be extracted in DIAsDEM must be appropriate to serve as elements of an XML DTD. Among other implications, discovering a concept that is pe-

culiar to a single text unit is not sufficient for our purposes, although it may perfectly reflect the corresponding content. In order to derive a DTD, we need to discover groups of text units that share semantic concepts. Moreover, we concentrate on domain-specific texts that significantly differ from average texts with respect to word frequency statistics. These collections can hardly be processed using standard text mining software because the integration of domain knowledge is a prerequisite for successful knowledge discovery.

There are only a few research activities aiming at the transformation of texts into semantically annotated XML documents: Becker et al. introduce the search engine GETESS that supports query processing on texts by deriving and processing XML text abstracts [BBBD+2000]. These abstracts contain language-independent, content-weighted summaries of domain-specific texts. Decker et al. extract meta-data from Web documents using the ontology-based system ONTOBROKER [DEFS1999]. Maedche and Staab introduce an architecture for semi-automatically learning ontologies from Web documents [MaSt2001]. In DIAsDEM, we do not separate meta-data from original texts but rather provide a semantic annotation, keeping the texts intact for later processing. Given the aforementioned linguistic particularities of the application domains we investigate, a DTD characterizing the content of documents is more appropriate than inferences from their content.

In order to transform existing content into XML documents, Sengupta and Purao propose a method that infers DTDs by using already tagged documents as input [SePu2000]. In contrast, we propose a method that tags plain text documents and derives a DTD for them. Closer to our approach is the work of Lumera who uses keywords and rules to semi-automatically convert legacy data into XML documents [Lumera2000]. However, his approach relies on establishing a rule base that drives the conversion, while we use a KDD approach to reduce human effort.

Semi-structured data is another topic of related research within the database community [Buneman1997, AbBuSu2000]. A lot of effort has recently been put into methods inferring and representing structure in similar semi-structured documents [NeAbMo1997, WaLi2000, LaMaPo2000]. However, these approaches only derive a schema for a given set of semi-structured documents. In DIAsDEM, we have to simultaneously solve the problems of both semi-structuring text documents by semantic tagging and inferring an appropriately structured XML DTD that describes the related archive.

THE DIASDEM FRAMEWORK

In DIAsDEM, the notion of semantic tagging refers to the activity of annotating texts with domain-specific XML tags. Rather than classifying entire documents or tagging single terms, we aim at semantically tagging structural text units such as sentences or paragraphs. The semantics of text units are made explicit by XML tags that may contain additional at-

tributes describing named entities (e.g., persons, companies or amounts of money).

Our framework pursues two objectives for a given archive of text documents: All text documents should be semantically tagged and an appropriate, preliminary flat XML DTD should be derived for the archive. Semantic tagging in DIAsDEM is a two-phase process. We have designed a knowledge discovery in textual databases (KDT) process that constitutes the first phase in order to build clusters of semantically similar text units, to tag documents in XML according to the results and to derive an XML DTD describing the archive. The KDT process that was introduced in [GrSpWi2001, GrWiSp2001] results in a final set of clusters whose labels serve as XML tags and DTD elements. Huge amounts of new text documents from the same domain can be converted into XML documents in the second, batch-oriented and productive phase of the DIAsDEM framework. All text units contained in new documents are clustered by the previously built text unit clusterer and are subsequently annotated with the corresponding XML tags.

In DIAsDEM, we concentrate on the semantic tagging of similar text documents originating from a common domain. Nevertheless, the approach is appropriate for semantically tagging various kinds of archives such as public announcements of courts and administrative authorities, reports to shareholders and product descriptions on e-marketplaces.

The iterative and interactive KDT process that constitutes the first phase of the DIAsDEM framework is depicted in Figure 1. It is termed “iterative” because the clustering algorithm is invoked repeatedly. However, our notion of iterative clustering should not be confused with the fact that most clustering algorithms perform multiple passes over the data before converging. This process is also “interactive”, because a knowledge engineer is consulted for cluster evaluation and final cluster labeling decisions.

Besides the initial text documents to be tagged, the following domain knowledge constitutes input to our KDT process: A controlled vocabulary (e.g., thesaurus) containing a domain-specific taxonomy of terms and concepts, a preliminary reference schema (e.g., UML schema) of the domain and descriptions of specific named entities of importance, e.g. persons and companies. The reference schema reflects the semantics of named entities and relationships among them, as application experts initially conceive them. This schema serves as a reference for the DTD to be derived from discovered semantic tags, but there is no guarantee that the final DTD will be contained in or will contain this schema.

Our KDT process starts with a pre-processing phase: After setting the level of granularity by determining the size of text units, the Java- and Perl-based DIAsDEM Workbench performs basic NLP pre-processing such as tokenization, normalization and word stemming using TreeTagger [Schmid1994]. Instead of removing stop words, we establish a drastically reduced feature space by selecting a limited set of terms and concepts (i.e. text unit descriptors) from the controlled vocabulary and the reference schema. The knowledge en-

gineer currently chooses text unit descriptors, because they must reflect important concepts of the application domain. All text units are mapped onto Boolean vectors of this feature space. Thereafter, all Boolean text unit vectors are further processed by applying a standard IR weighting schema (TFxIDF). Additionally, named entities of interest are extracted from text units by a separate module of the DIAsDEM Workbench.

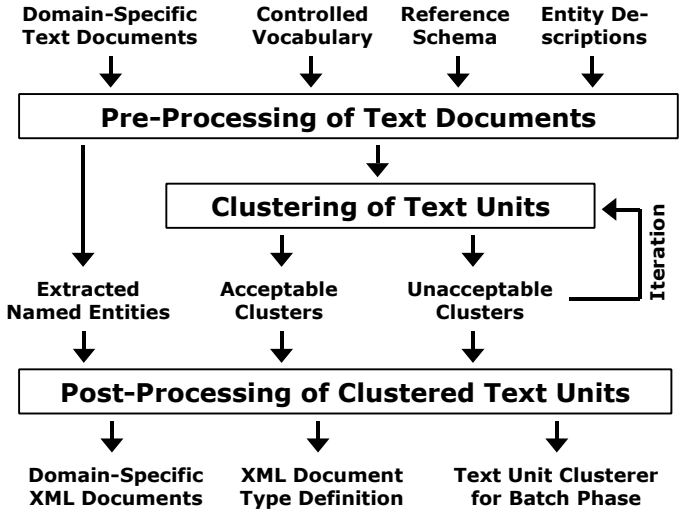

Figure 1 – Iterative and interactive KDT process of the DIAsDEM framework

In the pattern discovery phase, all text unit vectors contained in the initial archive are clustered based on similarity of their content. The objective is to discover dense and homogeneous text unit clusters. Clustering is performed in multiple iterations. Each iteration outputs a set of clusters, which the DIAsDEM Workbench partitions into qualitatively "acceptable" and "unacceptable" ones according to our cluster quality criteria. A cluster of text unit vectors is "acceptable", if and only if (i) its cardinality is large and the corresponding text units are (ii) homogeneous and (iii) can be semantically described by a small number of text unit descriptors. Members of "acceptable" clusters are subsequently removed from the dataset for later labeling, whereas the remaining text unit vectors are input data to the clustering algorithm in the next iteration. In each iteration, the cluster similarity threshold value is stepwise decreased such that "acceptable" clusters become progressively less specific in content. The KDT process is based on a plug-in concept that allows the execution of different clustering algorithms within the DIAsDEM Workbench.

The post-mining phase consists of a labeling step, in which "acceptable" clusters are semi-automatically assigned a label. Ultimately, the knowledge engineer determines cluster labels. However, the DIAsDEM Workbench performs both a pre-selection and a ranking of candidate cluster labels for the expert to choose from. All default cluster labels are derived from feature space dimensions (i.e. text unit descriptors) prevailing in the respectable

clusters. Cluster labels actually correspond to XML tags that are subsequently used to annotate cluster members. Finally, all original documents are tagged using valid XML tags. In addition, XML tags are enhanced by attributes reflecting previously extracted named entities and their values. The DIAsDEM Workbench finally derives a currently flat and unstructured XML DTD that coarsely describes the semantic structure of the resulting XML collection.

In order to evaluate the quality of our approach in absence of pre-tagged documents, we propose to draw a random sample of text units and ask a domain specialist to verify their annotations with respect to the following error types: Firstly, a text unit is annotated with a wrong XML tag, i.e. the tag does not properly reflect the content of the text unit (false positive). Secondly, a text unit is not annotated at all, although there exists an XML tag in the derived DTD reflecting the content of the text unit (false negative).

CASE STUDY: TAGGING PUBLIC TEXT ARCHIVES

DIAsDEM is a general-purpose framework whose workbench can be coupled with application-specific controlled vocabularies (e.g., thesauri) and named entity descriptions as well as various clustering algorithms.

The Application Domain

We applied our framework to text documents of the German Commercial Register. In Germany, each district court maintains a Commercial Register that contains important information about the companies in the court's district. According to law, many company activities like the establishment of branch offices, changes of share capital or mergers and acquisitions must be reported to the respective register.

Knowledge of Commercial Register entries is indispensable for business activities, as they have both a right-confirmation and a right-generating effect according to German law. For example, consider two companies preparing a contract, which must be signed by authorized physical persons. A person is an authorized representative of a company, if and only if there is an entry in the register granting authorization rights to this person and if there is no later entry, in which the authorization is revoked. Commercial Register entries are made available to the public, since up-to-date knowledge about a company's affairs is essential to its (prospective) stakeholders. The availability of Commercial Register documents on the Web has thus a large potential for focused information acquisition.

Indeed, due to the intense business demand for this type of commercial information, there are several information brokers offering both online and offline services to retrieve information from Commercial Registers. Current services only encompass SQL queries to access relational data and full-text queries to search the texts containing most of the information.

Enhancing textual entries with semantic meta-data would largely facilitate searching, browsing and querying. Tagged documents may also serve as pre-processed input to further KDT algorithms aimed at obtaining useful knowledge about companies. Therefore, we have chosen Commercial Register entries to evaluate our framework.

HRB 12990 **30.09.1999**	**Behrens & Klein Oberbausysteme GmbH** **(Seeblickstraße 26, 15758 Zernsdorf)**	**publiziert am** **09.10.1999**
Vertrieb und Entwicklung von Gleisoberbautechnik. Stammkapital: 50.000 DM. Gesellschaft mit beschränkter Haftung. Der Gesellschaftsvertrag ist am 02. Dezember 1994 abgeschlossen. Durch Beschluss der Gesellschafterversammlung vom 07. April 1999 ist der Sitz der Gesellschaft von Berlin nach Zernsdorf verlegt und der Gesellschaftsvertrag geändert in § 1 (Sitz). Ist nur ein Geschäftsführer bestellt, so vertritt er die Gesellschaft einzeln. Sind mehrere Geschäftsführer bestellt, so wird die Gesellschaft durch zwei Geschäftsführer oder durch einen Geschäftsführer in Gemeinschaft mit einem Prokuristen vertreten. Einzelvertretungsbefugnis kann erteilt werden. Hendrik Klein, 16.02.1967, Zernsdorf, und Klaus Behrens, 04.01.1958, Braunschweig, sind zu Geschäftsführern bestellt. Sie vertreten die Gesellschaft stets einzeln und sind befugt, Rechtsgeschäfte mit sich im eigenen Name oder als Vertreter eines Dritten abzuschließen. Nicht eingetragen: Die Bekanntmachungen der Gesellschaft erfolgen im Bundesanzeiger.		

Table 1 – Example entry in a German Commercial Register

Table 1 contains a German Commercial Register entry. As this example illustrates, each entry consists of a structured part and an unstructured text. 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 and can afterwards be stored in a relational DBMS. The unstructured section of each entry contains the registered text as recorded by the court's clerks. Three main categories of Commercial Register entries can be distinguished: foundation entries of new companies, update entries (e.g., changes in the managerial head of a company) and entries announcing that a company closes.

We have conceptually modeled the application domain using UML class diagrams. This conceptual model serves as a reference, against which the derived DTD can be matched, and supports the creation of an application-specific thesaurus. In Figure 2, the base information about this domain is depicted. We stress the fact that the information on one company is dispersed along as many courts as are the districts in which the company maintains branch offices. Figure 3 depicts a simplified taxonomy of German company types. Each type is regulated by special law and thus requires the registration of different data.

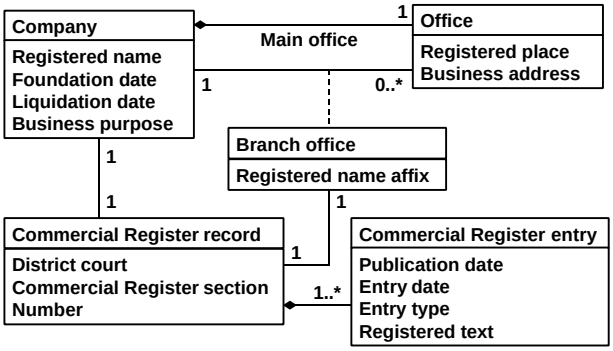

Figure 2 – Fundamental information about the domain (UML class diagram)

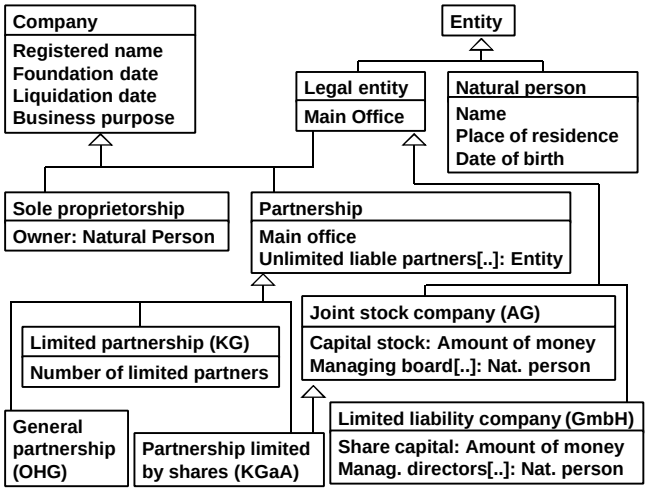

Figure 3 – Simplified taxonomy of German company types (UML class diagram)

Building the Text Unit Clusterer

In the first phase of our framework, we create a set of clusters (i.e. the "text unit clusterer") of similar text units. To build this clusterer, we have used 1,145 documents published by the district court of Potsdam. These documents are foundation entries of new companies for the year 1999. The conceptual model depicted in Figures 2 and 3 formed the basis for specifying a domain-specific controlled vocabulary. The DIAsDEM Workbench supports thesaurus generation by word frequency statistics and a thesaurus editor. Using it, we have built a domain-specific thesaurus containing a hierarchy of 70 descriptors and 109 non-descriptors pointing to valid descriptors. The final feature space consists of 85 text unit descriptors, after adding terms known to be of importance in this particular domain.

We have partitioned the documents into text units, whereby the text unit was set to a sentence. Afterwards, the multilingual part–of–speech tagger TreeTagger [Schmid1994] was

employed to determine lemma forms of all words. The TreeTagger reduced the number of unique word forms from 10,613 to approx. 5,400. Our Perl- and Java-based DIAsDEM Workbench was employed to identify instances of named entities such as “Person”, “Company”, “Date” and “AmountOfMoney” and to map all text units into text unit vectors of the final feature space that consists of 85 text unit descriptors. Figure 4 exemplary illustrates all steps performed during the pre-processing phase of our proposed KDT process.

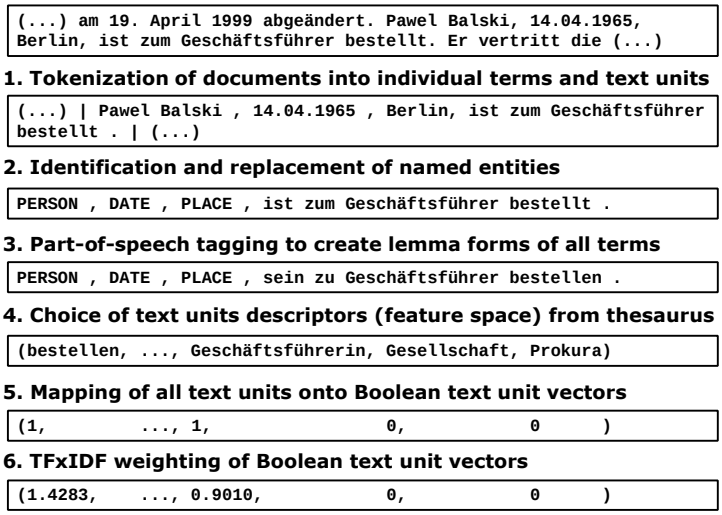

Figure 4 – Exemplary outline of the KDT pre-processing phase

The IBM DB2 Intelligent Miner for Data was afterwards applied to detect groups of semantically similar text unit vectors by explorative pattern discovery [IBM2001]. In particular, we used its demographic clustering function whose objective is the maximization of the value of Condorcet’s criterion [Michaud1997]. This criterion is the difference between 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 maximum number of clusters to be generated can by limited but is afterwards automatically determined by the miner. The similarity threshold is another parameter of the clustering algorithm that adjusts the assignment of text unit vectors to clusters. The collection of text units from the Potsdam collection has been clustered iteratively. In Table 2, we summarize the size of the dataset and the parameter settings of each iteration.

After each iteration, the DIAsDEM Workbench displays the content of qualitatively “acceptable” clusters. For each cluster, the knowledge engineer also obtains ranked default cluster descriptions, from which the expert creates an appropriate cluster label. The remaining “unacceptable” clusters are input to the next KDT iteration. After three iterations, 73 homogeneous clusters were identified. They represent approx. 85% of all text units in the collection, as summarized in Table 2.

Clustering iteration of KDT process	*1*	*2*	*3*
Number of input text units	10,785	1,818	1,648
Similarity threshold	0.95	0.90	0.80
Maximum number of clusters	200	200	200
Visualization threshold (cluster size)	10	5	3
Number of output clusters	122	121	67
Global Condorcet value	0.8090	0.9147	0.8176
Number of acceptable clusters	42	12	19
Text units in acceptable clusters	8,969	168	74

Table 2 – Input, parameters and result statistics of each KDT iteration

In the tagging phase, the semantic labels of "acceptable" clusters were used to annotate their member sentences with corresponding XML tags. Sentences not clustered at all or sentences belonging to unlabeled clusters, i.e. to qualitatively "unacceptable" clusters, remained un-tagged. Additionally, XML tags were extended with attributes corresponding to previously extracted named entities and their values. Table 3 illustrates the semantically tagged document of Table 1.

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

<CommercialRegisterEntry> <BusinessPurpose> Vertrieb und Entwicklung von
Gleisoberbautechnik. </BusinessPurpose> <ShareCapital AmoutOfMoney="50000 DM">
Stammkapital: 50.000 DM. </ShareCapital> <LimitedLiabilityCompany> Gesellschaft
mit beschränkter Haftung. </LimitedLiabilityCompany> <ConclusionArticles Date=
"02.12.1994"> Der Gesellschaftsvertrag ist am 02. Dezember 1994 abgeschlossen.
<ConclusionArticles> <ModificationArticles_MainOffice Date="07.04.1999"
Paragraph="§ 1 (Sitz)"> Durch Beschluss der Gesellschafterversammlung vom 07.
April 1999 ist der Sitz der Gesellschaft von Berlin nach Zernsdorf verlegt und der
Gesellschaftsvertrag geändert in § 1 (Sitz). </ModificationArticles_MainOffice>
(...) Einzelvertretungsbefugnis kann erteilt werden. <AppointmentManagingDirector
Person="Klein; Hendrik; Zernsdorf; 16.02.1967 && Behrens; Klaus; Braunschweig;
04.01.1958"> Hendrik Klein, 16.02.1967, Zernsdorf, und Klaus Behrens, 04.01.1958,
Braunschweig, sind zu Geschäftsführern bestellt. </AppointmentManagingDirector>
(...) <PublicationMedia> Nicht eingetragen: Die Bekanntmachungen der Gesellschaft
erfolgen im Bundesanzeiger. </PublicationMedia> </CommercialRegisterEntry>
```

Table 3 – XML document containing an annotated Commercial Register entry

The first sentence of the document in Table 3 is tagged as one referring to the business purpose of the new company. The second sentence refers to the share capital of the company. This sentence contains a named entity, i.e. the amount of money invested in the company. Accordingly, its XML tag is extended to accommodate the entity name "AmountOfMoney" and its value. One tag refers to the managing director appointed for the company and thus contains the named entity "Person". Its value reflects the way persons are identified in many entries, i.e. by specifying surname, forename, current domicile and date of birth. Table 4 contains an excerpt of the flat, unstructured XML DTD that was automatically derived. It coarsely describes the semantic structure of the resulting XML collection. Currently, named entities are not fully evaluated as attributes of XML tags.

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

<!ELEMENT CommercialRegisterEntry ( #PCDATA | BusinessPurpose |
ResolutionByShareholders_ShareCapital | ShareCapital | ModificationMainOffice |
FullyLiablePartner | AppointmentManagingDirector | GeneralPartnership |
InitialShareholders | NonCashCapitalContribution | LimitedLiabilityCompany |
ConclusionArticles | DivisionCapitalStock | (...) | FoundationPartnership )* >

<!ELEMENT BusinessPurpose (#PCDATA)>
<!ELEMENT ShareCapital (#PCDATA)> (...)
<!ELEMENT FoundationPartnership (#PCDATA)>

<!ATTLIST ShareCapital AmountOfMoney CDATA #IMPLIED> (...)
<!ATTLIST AppointmentManagingDirector Person CDATA #IMPLIED> (...)
<!ATTLIST ConclusionArticles Date CDATA #IMPLIED> (...)
```

Table 4 – Preliminary flat, unstructured XML DTD of Commercial Register entries

In our application domain, there are no pre-classified documents, upon which the effectiveness of the DIAsDEM Workbench could be measured. Instead, we have drawn a random sample containing approx. 5% of 1,145 text units and had them inspected by a domain expert to detect tagging errors. The results of our evaluation are shown in Table 5 whose second column summarizes the evaluation results concerning the effectiveness of the induced text unit clusterer for the Potsdam archive. It can be stated that the percentage of false positives is very low: if a text unit is tagged, its tag is most likely to be correct. Hence, a service processing XML documents tagged with our mechanism can rely on the correctness of the tags.

District court of collection	*Potsdam (training)*	*Berlin (application)*
Number of input text documents	1,145	3,954
Number of input text units	10,785	36,344
Number of text units in 5% sample	533	1,845
False positive error rate in 5% sample	0.4%	0.7%
False negative error rate in 5% sample	3.6%	13.3%
Overall error rate in 5% sample	4.0%	14.0%

Table 5 – Evaluation statistics of both training and application collection

The percentage of false negatives is higher, indicating that some text units were not placed in the cluster they semantically belonged to. Our preliminary explanation for the comparatively high rate of false negatives is that these text units were characterized by words that were not included in the feature space. The reader may recall that there was no thesaurus available for this case study, so that one had to be built from word statistics. A thesaurus contains several concepts, each of them expressed with many alternative words: if some of these alternatives are less frequent than others, they may be ignored when building the thesaurus and deriving a feature space for it. This has the effect that text units containing these infrequent words are mapped onto vectors of poor quality.

The last line of the second column of Table 5 computes the overall error rate in the sample. With 0.95 confidence, this error is in the interval [2.6%, 6.0%], which is a very promising result. If our explanation for the false negatives is correct, extending the feature space to take some less frequent synonyms into account will reduce the error.

Application of the Text Unit Clusterer

A company with multiple branch offices may have entries in Commercial Registers of several courts. Thus, searching for data related to a company should span several document collections. This implies that semantic tags must be derived from all registers in Germany. Since the information to be reported to Commercial Registers is determined by federal law, the semantics of documents can be expected to be independent of the court where the data are registered. Thus, a set of semantic tags and the corresponding DTD derived from the Commercial Register of one court should be appropriate for all German registers. We have thus applied the text unit clusterer built from the Potsdam texts to a document collection from the Commercial Register in Berlin. This collection consists of 3,954 entries on company foundations from February to September 1999 and contains 36,344 text units.

Column three of Table 5 summarizes the evaluation statistics of applying the previously derived text unit clusterer to the Berlin collection. Again, an expert has manually inspected a 5% random sample of text units. The overall error rate of 36,344 Berlin text units is in the interval [12.5%, 15.6%] with 0.95 confidence. This is significantly higher than the error rate of the Potsdam collection. The difference is mainly caused by the dramatic increase of the false negative rate, while the rate of false positives is still less than 1%. This indicates that text units that could be tagged were annotated correctly with high confidence. However, much less text units could be assigned to the appropriate cluster. A closer inspection of the Berlin collection and cross-checking with the Potsdam archive revealed that there are many differences in formulating the same concept. Apparently, the clerks who type registered texts follow conventions that seem to differ from court to court. Hence, an interpretation would be that the feature space for the Berlin collection should be slightly different from that for Potsdam. Moreover, the syntactic conventions are not the same, to the effect that information put into two sentences in one collection appears within a single sentence in the other. This may also be a cause for false negatives, since the text unit was defined to be a sentence.

CONCLUSION

In this paper, we have summarized the DIAsDEM framework for semi-automated semantic tagging of domain-specific texts. This framework has been successfully applied to collections of the publicly available German Commercial Register. However, Commercial Register entries are composed of rather regular and antiquated German language, which

might contribute to the low overall error rates in the case study. Therefore, we are currently working on a different application domain characterized by a greater linguistic diversity: Ad hoc news is issued by publicly quoted companies and contains information about current developments that potentially influence share prices. Both stakeholders and public authorities pursuing investor protection, market transparency and market integrity have a particular interest in this public source of information. This latest case study is not finished yet. However, the preliminary results of evaluating the quality are once again promising.

Currently, we are developing the prototype of a Web-based information system that enables companies and interested citizens to query semantically annotated Commercial Register documents. In contrast to conventional full-text search, this IT system will also support structure-based queries exploiting XML tags and their attributes. To this end, we will evaluate XML query languages supporting both structure-based and content-based queries.

Of course, many open research issues remain: First of all, the method of deriving XML DTDs should be refined to reflect the complexity of documents. We are aiming at structuring the preliminary and flat XML DTD by discovering ordered and nested tags. Firstly, we are going to employ sequence mining techniques in order to discover the most likely ordering of DTD elements. Secondly, we plan to discover nested XML tags by employing a hierarchical clustering algorithm within the DIAsDEM Workbench. The resulting hierarchy of clusters and sub-clusters might correspond to a hierarchy of nested XML tags that must of course be approved by the knowledge engineer. Varying the level of granularity of text units (e.g., from noun groups to sentences and further to n consecutive sentences) within the KDT process might also be beneficial to discover nested DTD elements. Since all tags are discovered by data mining techniques, we will introduce the notion of a probabilistic DTD to cater for inevitable tagging errors. If applicable, attributes of XML tags will be assigned a semantic name as well.

Particularly in Europe, the semantic tagging of multilingual texts is another challenge. Provided with a multilingual thesaurus and a language identifier for each document, our framework should be general enough to be applied to this type of archives. Ultimately, our objective is the integration of archives described by probabilistic XML DTDs with related structured and semi-structured data sources into a homogeneous information system.

ACKNOWLEDGMENTS

We thank the German Research Society DFG for funding the project DIAsDEM (DFG grant no. SP 572/4-1), the Bundesanzeiger Verlagsgesellschaft mbH for providing data and our project collaborators Evguenia Altareva and Stefan Conrad for helpful discussions. The IBM Intelligent Miner for Data is kindly provided by IBM in terms of the IBM DB2 Scholars Program.

REFERENCES

[AbBuSu2000] S. Abiteboul, P. Buneman, and D. Suciu. *Data on the Web: From Relations to Semistructured Data and XML*. Morgan Kaufman Publishers, San Francisco, 2000.

[BBBD+2000] 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.

[Buneman1997] 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.

[DEFS1999] S. Decker, M. Erdmann, D. Fensel, and R. Studer. ONTOBROKER: Ontology based access to distributed and semi-structured information. In *Proceedings of the TC2/WG 2.6 8th Working Conference on Database Semantics*, Rotorua, New Zealand, 1999.

[FaPiSm1996] 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.

[FFKL+1998] 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.

[GrSpWi2001] H. Graubitz, and M. Spiliopoulou, and K. Winkler. The DIAsDEM framework for converting domain-specific texts into XML documents with data mining techniques. In *Proceedings of the First IEEE International Conference on Data Mining*, San Jose, CA, USA, November/December 2001. To appear.

[GrWiSp2001] H. Graubitz, K. Winkler, and M. Spiliopoulou. Semantic tagging of domain-specific text documents with DIAsDEM. In *Proceeding of the 1st International Workshop on Databases, Documents, and Information Fusion (DBFusion 2001)*, pages 61–72, Magdeburg, Germany, May 2001.

[IBM2001] IBM DB2 Intelligent Miner for Data. http://www.ibm.com/software/data/iminer. Accessed 2001-11-05.

[LaMaPo2000] P. A. Laur, F. Masseglia, and P. Poncelet. Schema mining: Finding regularity among semistructured data. In *Principles of Data Mining and Knowledge Discovery: 4th European Conference, PKDD 2000*, pages 498–503, Lyon, France, September 2000.

[LoWiOl2000] 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.

[Lumera2000] J. Lumera. Große Mengen an Altdaten stehen XML-Umstieg im Weg. *Computerwoche*, 27(16):52–53, 2000.

[MaSt2001] A. Maedche and S. Staab. Learning ontologies for the Semantic Web. In *Proceedings of the 2nd International Workshop on the Semantic Web - SemWeb'2001*, Hongkong, China, May 2001.

[Michaud1997] P. Michaud. Clustering techniques. *Future Generation Computer Systems*, 13(2-3):135–147, 1997.

[MiFi1995] 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.

[NaMo2000] 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.

[NeAbMo1997] S. Nestrov, S. Abiteboul, and R. Motwani. Inferring structure in semi-structured data. *SIGMOD Record*, 26(4):39–43, 1997.

[Schmid1994] 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.

[SePu2000] 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.

[WaLi2000] 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.