Semantic Tagging of Domain-Specific Text Archives with Data Mining Techniques

Dipl.-Kfm. Karsten Winkler[*]

Leipzig Graduate School of Management (HHL)
Department of Information Systems and E-Business
Jahnallee 59, D-04109 Leipzig, Germany
Phone: +49 341 9851 760, Fax: +49 341 9851 764
E-Mail: kwinkler@ebusiness.hhl.de

Advisor: Prof. Dr. Myra Spiliopoulou

Leipzig Graduate School of Management (HHL)
Department of Information Systems and E-Business
Jahnallee 59, D-04109 Leipzig, Germany
Phone: +49 341 9851 760, Fax: +49 341 9851 764
E-Mail: myra@ebusiness.hhl.de

[*] The author has been funded by the German Research Society (DFG grants no. SP 572/4-1 and SP 572/4-3) within the research project DIAsDEM. This German acronym stands for "Data Integration of Legacy Data and Semi-Structured Documents Using Data Mining Techniques".

Abstract

Organizations accumulate large and continuously growing volumes of text documents. Unfortunately, most textual 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 is capable of solving these challenges by providing a solid basis for information systems that provide effective and efficient access to textual data. In contrast to plain texts, semantically tagged XML documents along with appropriate query languages facilitate searching and browsing, knowledge management and information integration. However, transforming textual data into semantically annotated XML documents should be largely automated to minimize costly human efforts.

In this doctoral work, the DIAsDEM framework for semi-automated semantic tagging of domain-specific text documents is thus being developed, implemented and evaluated using real-world datasets. This new framework includes a complex knowledge discovery process. It exploits the fact that many archives contain documents sharing an inherent though undocumented structure, although they are composed of unstructured texts. In order to semantically annotate documents, this inherent structure should be made explicit by XML tags.

The knowledge discovery process groups structural text units (e.g., sentences) based on similarity of their contents. A clustering algorithm is executed iteratively, whereas each iteration outputs a set of acceptable text unit clusters according to specific cluster quality criteria. Acceptable clusters are labeled with default semantic names which are refined by experts. Thereafter, cluster labels serve as semantic XML tags for the corresponding text units. XML tags are enhanced with attributes containing previously extracted named entities (e.g., names of companies). Finally, an archive-specific and appropriately structured XML document type definition (DTD) is derived that contains meta-data describing the semantics of the archive. This doctoral research also includes commercially relevant case studies to evaluate the tagging quality and to refine the pursued approach. For example, Commercial Register entries and ad hoc news of publicly traded companies are semantically annotated.

1 Introduction

Apparently, both commercial and non-profit organizations create and store huge volumes of data [16]. Instead of utilizing these large datasets to efficiently attain their respective objectives, there exists a wide-spread fear of "drowning" in one's own data [6]. Hence, knowledge discovery in databases (KDD) has become an active research field for the past years. Aiming at the extraction of new, non-trivial and actionable knowledge from data, knowledge discovery in databases combines various methods from statistics, machine learning, artificial intelligence and database research in a unifying framework [8].

With respect to the degree of internal structure, structured such as relational data can be distinguished from semi-structured (e.g., HTML files) and unstructured documents (e.g., texts). Currently, up to 80% of a company's information is contained in unstructured text documents [28]. Hence, KDD techniques have been developed to classify texts or to discover similar documents for the previous years as well. Analogously to data warehousing and data mining, *document warehousing* and *text mining* are emerging terms that denote techniques for capturing and utilizing the flood of textual information for decision making [27].

Organizations tend to have large, domain-specific and electronically accessible text archives of rather homogeneous documents at their disposal. They might contain for example project reports or various types of internal memos. In contrast, only few organizations currently utilize these textual archives to create additional value for their stakeholders [5]. In most cases, only conventional and thus limited full-text search is employed to retrieve relevant information. Instead of or in addition to using full-text search, exploiting existing semantic structure and application-specific objects (e.g., customers, companies or products) in queries often offers large benefits in terms of retrieval performance [2, pp. 106–113] [10].

Innovative companies gradually realize that a purposeful management of both explicit and implicit knowledge provides vast opportunities for creating sustainable, knowledge-based competitive advantages. Undoubtedly, text archives are one major source of explicit organizational knowledge. In this context, the fine-grained semantic tagging of text archives to explicate inherent structure is an important, but rather neglected means of extracting new, non-trivial and actionable knowledge from and about textual resources.

2 Research Motivation and Objectives

The semantic annotation of text archives using the Extensible Markup Language XML results in application-specific, semantic meta-data in the form of XML tags and an archive-specific XML document type definition (DTD). Semantic meta-data can be utilized to facilitate for example knowledge management and information integration. Appropriate XML query languages could for example be employed to submit both content- and structure-based queries against XML archives. However, two main problems must be solved to semi-automatically create text annotations: Firstly, an appropriately structured, semantic DTD should be derived for each textual archive. Secondly, all text documents contained in an archive should be semantically tagged according to the previously derived DTD.

Currently, most methods of knowledge discovery in textual databases (KDT) either analyze contents at the document level or focus on the term level by applying natural language processing techniques. These two approaches exhibit limitations in the context of knowledge discovery in domain-specific archives containing rather homogeneous texts, because the coarse document content is already known. On the other hand, linguistically complex, domain-specific documents often differ from average language usage with respect to syntax and vocabulary. Therefore, domain knowledge should be incorporated into the KDD process. Moreover, the important and discriminating information is contained in fine-grained, structural text components. In contrast to current approaches, this doctoral research aims at designing a process-oriented methodology that enables knowledge discovery at the level of structural text units (e.g., sentences or paragraphs).

Hence, the primary objectives of this doctoral work are the design of a conceptual framework for semantically semi-structuring large, domain-specific archives of homogeneous text documents as well as the development of a research prototype and its evaluation in real-world case studies. The notion of “semi-structuring” refers (i) to the derivation of a semantic XML DTD that describes the inherent, though undocumented structure of the corresponding archive and (ii) to the subsequent markup of texts according to the derived DTD. In order to attain these objectives, semantically similar text units must be discovered, semantically labeled and aggregated into an appropriately structured DTD, whereby the necessary human efforts should be minimized. Additionally, developing commercial usage scenarios and estimating the market volume for the corresponding applications constitute secondary objectives of this doctoral work.

3 Literature Review

For this doctoral research, three major fields of related work must be considered: Knowledge discovery in textual databases, research aimed at transforming unstructured texts into semantically annotated or semi-structured documents as well as schema discovery in collections of similar semi-structured documents.

Concerning related knowledge discovery work, Nahn and Mooney propose the combination of methods from KDD and information extraction to perform text mining tasks [19]. 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 [9]. 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. The DIAsDEM framework adopts the idea of representing texts by terms and concepts using the vector model proposed by Salton et al. [24]. 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 [14]. Similarly to our framework, the authors suggest to exploit existing controlled vocabularies such as thesauri for concept extraction. Mikheev and Finch describe a workbench to acquire domain knowledge from texts [17]. Similarly to the prototypically implemented DIAsDEM Workbench, their approach combines methods from different fields of research in a unifying framework.

The DIAsDEM approach shares with this research thread the objective of extracting semantic concepts from texts. However, concepts to be extracted in this doctoral work must be appropriate to serve as elements of the XML DTD to be derived. Among other implications, discovering a concept that is peculiar to a single text unit is not sufficient for our purposes, although it may perfectly reflect the corresponding content. In order to derive a document type definition, we need to discover groups of text units that share some semantic concepts. Moreover, we concentrate on domain-specific texts, which 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 relevant domain knowledge is a prerequisite for successful knowledge discovery.

Currently, there are only a few research activities aiming at the transformation of texts into semantically annotated XML documents: Bruder et al. introduce the search engine GETESS that supports query processing on texts by creating and processing XML text abstracts [3]. These abstracts contain language-independent, content-weighted summaries of domain-specific texts. In DIAsDEM, we do not separate meta-data from original texts but rather provide a semantic annotation, keeping the texts intact for later processing or visualization. Given the aforementioned linguistic particularities of the application domains we investigate, a DTD characterizing the contents of the entire archive is more appropriate than inferences on the contents of individual documents. Additionally, the GETESS approach requires an a priori given DTD that corresponds to a domain-specific ontology. Erdmann et al. introduce a system that supports a semi-automated and ontology-based semantic annotation of Web pages [7]. The authors associate previously extracted [23] text fragments (mostly named entities) with concepts of an a priori given ontology. In contrast, this doctoral work aims at deriving an XML DTD from unstructured text documents.

In order to transform existing contents into XML documents, Sengupta and Purao propose a method that infers DTDs by using already tagged documents as input [26]. In contrast, we propose a method that tags plain text documents and derives a DTD for them. Moore and Berman present a technique to convert textual pathology reports into XML documents [18]. In contrast to this work, the authors neither derive an XML DTD nor apply a knowledge discovery methodology. They rather employ natural language processing techniques and a medical thesaurus to map terms and noun groups onto medical concepts. Thereafter, medical concepts serve as XML tags that semantically annotate the corresponding terms in pathology reports. Closer 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 establishing a rule base that drives the conversion, while we use a KDD methodology that reduces necessary human intervention.

Semi-structured data is another topic of related research within the database community [1, 4]. A lot of effort has recently been put into methods inferring and representing structure in similar semi-structured documents [13, 20, 22, 29]. 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 archive.

The related literature can be summarized as follows: The author is not aware of any scientific projects or commercial applications that aim at deriving an appropriately structured, semantic XML document type definition for domain-specific text archives by employing a knowledge discovery methodology.

4 Research Methodology and Contribution

Knowledge discovery in texts by creating semantic markup and deriving an XML DTD is an enabling methodology to transform textual contents into valuable assets. This doctoral research is conducted by applying a multimethodological approach to research as introduced by Nunamaker et al. [21]. The authors describe and defend the use of systems development as a methodology in information systems research. Their integrated research approach consists of theory building, systems development, observation and experimentation.

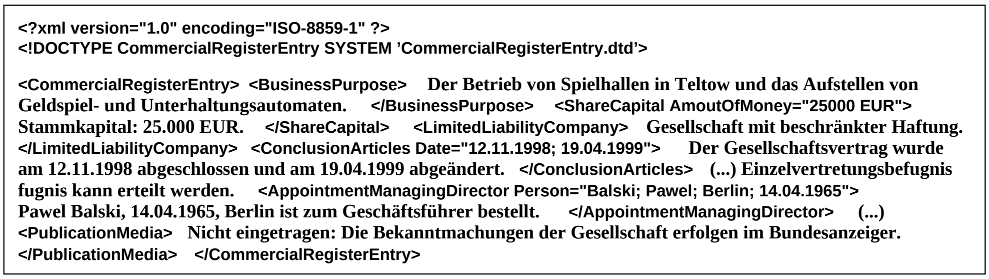
Consequently, this doctoral research employs the process for systems development research as described by Nunamaker et al. [21]. Concerning theory building, the main contribution of this doctoral work is the construction of a new conceptual framework that addresses the research question of semantic annotation of domain-specific text archives with data mining techniques. The systems development phase consists of designing the architecture of and building the prototype of the DIAsDEM Workbench. Furthermore, the DIAsDEM Workbench is evaluated in real-world case studies. Results from this observation and experimentation phase are used to stepwisely refine both the framework and the prototype.

The next subsection briefly summarizes the proposed approach to attain the primary research objectives. The results achieved so far comprise the design of the DIAsDEM framework for semantic tagging and for deriving a preliminary XML DTD as well as the implementation of the DIAsDEM Workbench as a research prototype that supports the entire framework. Additionally, we successfully finished a first case study to evaluate the quality of applying the framework to real-world text archives. Thereafter, both the current and remaining future doctoral research is summarized in a brief subsection.

4.1 The DIAsDEM Framework

In this doctoral work, the notion of semantic tagging refers to the activity of annotating texts with domain-specific XML tags that might contain additional attributes. Rather than classifying entire documents or tagging single terms, we aim at semantically tagging structural text units such as sentences or paragraphs. Figure 1 illustrates this concept of semantic tagging, whereas each sentence of this German Commercial Register entry is a text unit. In this example, the semantics of most sentences are made explicit by XML tags that partly contain additional attributes describing extracted named entities (e.g., names of persons and amounts of money). The XML document depicted in Figure 1 was created by applying the DIAsDEM framework to a collection of 1,145 textual Commercial Register entries containing 10,785 text units. This collection includes all entries related to foundations of companies in the district of the German city Potsdam in 1999. In Germany, companies are obliged by law to submit various information about business affairs to local Commercial Registers. Although Commercial Registers are an important source of information in daily business transactions, their textual contents can only be searched using full-text queries at the moment. Hence, semantically semi-structuring these textual archives provides the basis for information integration and creation of value-adding services related to information brokerage.

The framework pursues two objectives for an archive of text documents: All text documents should be semantically tagged and an appropriate, preliminary flat XML DTD should be derived for the archive. As illustrated in Figure 2, 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



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

<CommercialRegisterEntry>  <BusinessPurpose>    Der Betrieb von Spielhallen in Teltow und das Aufstellen von
Geldspiel- und Unterhaltungsautomaten.    </BusinessPurpose>    <ShareCapital AmoutOfMoney="25000 EUR">
Stammkapital: 25.000 EUR.    </ShareCapital>    <LimitedLiabilityCompany>    Gesellschaft mit beschränkter Haftung.
</LimitedLiabilityCompany>   <ConclusionArticles Date="12.11.1998; 19.04.1999">      Der Gesellschaftsvertrag wurde
am 12.11.1998 abgeschlossen und am 19.04.1999 abgeändert.   </ConclusionArticles>    (...) Einzelvertretungsbefugnis
fugnis kann erteilt werden.     <AppointmentManagingDirector Person="Balski; Pawel; Berlin; 14.04.1965">
Pawel Balski, 14.04.1965, Berlin ist zum Geschäftsführer bestellt.      </AppointmentManagingDirector>    (...)
<PublicationMedia>   Nicht eingetragen: Die Bekanntmachungen der Gesellschaft erfolgen im Bundesanzeiger.
</PublicationMedia>    </CommercialRegisterEntry>
```

Figure 1: XML document containing an annotated Commercial Register entry

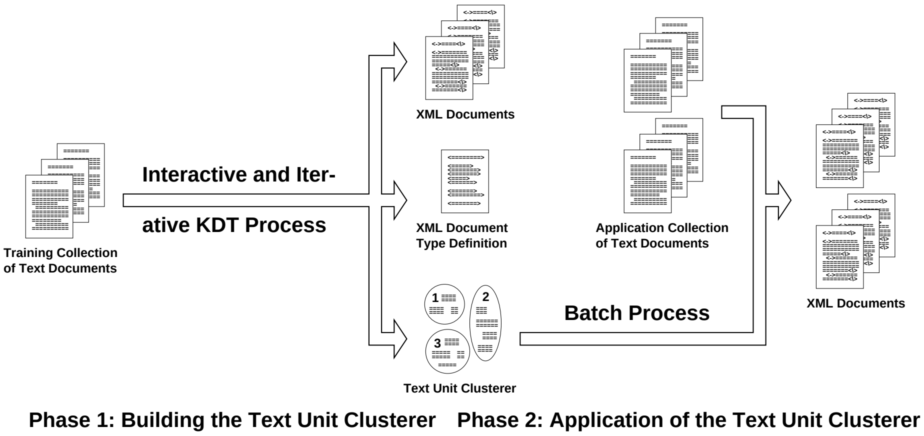

Training Collection of Text Documents
Interactive and Iterative KDT Process
XML Documents
XML Document Type Definition
Text Unit Clusterer
1
2
3
Application Collection of Text Documents
Batch Process
XML Documents
Phase 1: Building the Text Unit Clusterer
Phase 2: Application of the Text Unit Clusterer

Figure 2: Outline of the two-phase DIAsDEM framework

units, to tag documents in XML according to the results and to derive an XML DTD describing the archive. The KDT process [11, 12] results in a final set of clusters whose labels serve as XML tags and DTD elements. Huge amounts of new documents 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 tagged with the corresponding cluster labels.

This doctoral work focuses on the semantic tagging of similar text documents originating from a common domain. Nevertheless, the DIAsDEM approach is appropriate for semantically tagging various kinds of archives such as public announcements of courts and administrative authorities, quarterly and annual reports to shareholders, collections of company and industry news, textual patient records in health care applications as well as product and service descriptions published on electronic marketplaces.

In the remainder of this subsection, we briefly introduce the first phase of the DIAsDEM framework whose iterative and interactive KDT process is depicted in Figure 3. This process is termed "iterative" because the clustering algorithm is invoked repeatedly. 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",

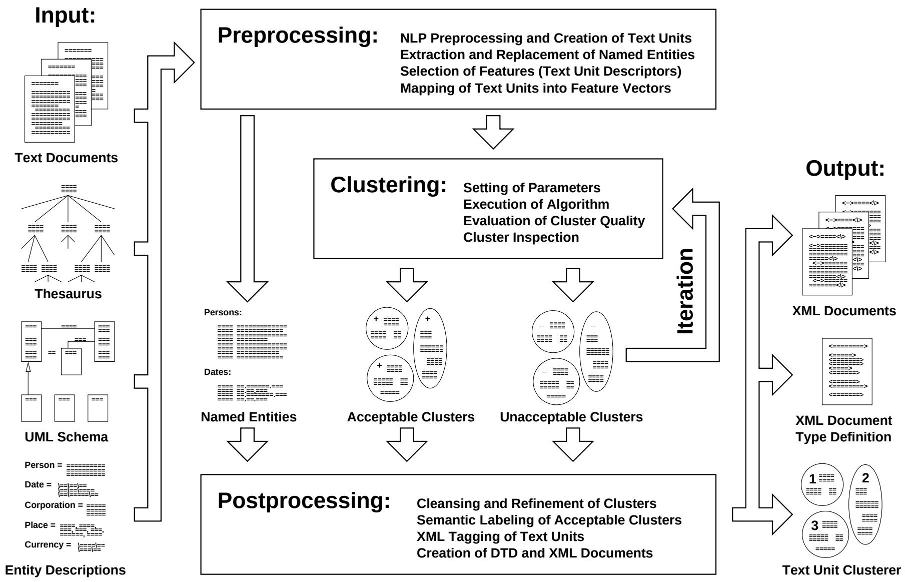

Input:
Text Documents
Thesaurus
UML Schema
Person =
Date =
Corporation =
Place =
Currency =
Entity Descriptions
Preprocessing:
NLP Preprocessing and Creation of Text Units
Extraction and Replacement of Named Entities
Selection of Features (Text Unit Descriptors)
Mapping of Text Units into Feature Vectors
Clustering:
Setting of Parameters
Execution of Algorithm
Evaluation of Cluster Quality
Cluster Inspection
Iteration
Persons:
Dates:
Named Entities
Acceptable Clusters
Unacceptable Clusters
Postprocessing:
Cleansing and Refinement of Clusters
Semantic Labeling of Acceptable Clusters
XML Tagging of Text Units
Creation of DTD and XML Documents
Output:
XML Documents
XML Document Type Definition
Text Unit Clusterer
1
2
3

Figure 3: Iterative and interactive KDT process of the DIAsDEM framework

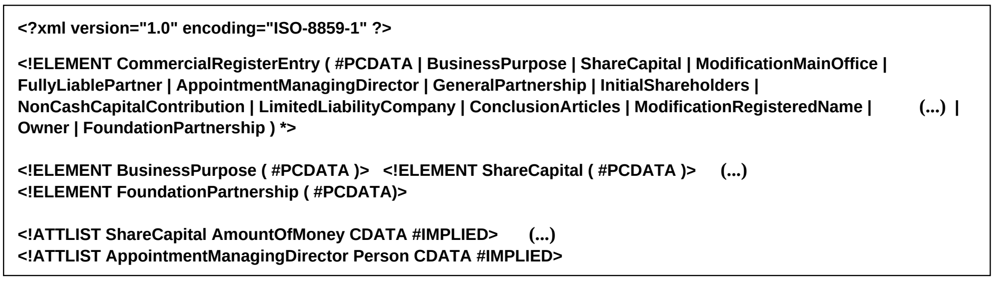
because a knowledge engineer is consulted for cluster evaluation and final cluster labeling decisions at the end of each iteration.

Besides the initial text documents to be tagged, the following domain knowledge constitutes input to our knowledge discovery process: A thesaurus containing a domain-specific taxonomy of terms and concepts, a preliminary UML schema of the domain and descriptions of specific named entities of importance, e.g. persons and companies. The UML schema reflects the semantics of named entities and relationships among them, as they are initially conceived by application experts. This schema might serve as a reference for the DTD to be derived from discovered semantic tags. However, there is no guarantee that the final document type definition will be contained in or will contain this schema.

Similarly to a conventional KDD process, our process starts with a preprocessing phase: After setting the level of granularity by determining the size of text units to be tagged, the Java- and Perl-based DIAsDEM Workbench performs basic NLP preprocessing such as tokenization, normalization and word stemming using TreeTagger [25]. Instead of removing stop words, we establish a drastically reduced feature space by selecting a limited set of terms and concepts (so-called text unit descriptors) from the thesaurus and the UML schema. Text unit descriptors are currently chosen by the knowledge engineer, because they should reflect important concepts of the application domain. All text units are mapped onto Boolean vectors of this feature space. Thereafter, each Boolean text unit vector is further processed by applying a standard information retrieval weighting schema, i.e. TF-IDF [24]. Additionally, named entities of interest are extracted from text units by the DIAsDEM Workbench. In our case study, we created a concise thesaurus and selected 70 relevant descriptors and 109 non-descriptors pointing to descriptors. The final feature space consisted of 85 text unit descriptors, after adding terms known to be of importance in this domain.

In the pattern discovery phase, all text unit vectors contained in the initial archive are clustered based on similarity of their contents. 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 ”acceptable” and ”unacceptable” ones according to our quality criteria. A cluster of text unit vectors is qualitatively ”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 stepwisely 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. In the case study, we employed the demographic clustering function included in the IBM Intelligent Miner for Data that maximizes the value of Condorcet’s criterion. After three iterations, the DIAsDEM Workbench discovered altogether 73 “acceptable” clusters containing approx. 85% of text units.

The postmining phase consists of a labeling step, in which “acceptable” clusters are semi-



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

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

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

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

Figure 4: Preliminary flat, unstructured XML DTD of Commercial Register entries

automatically assigned a label. Ultimately, cluster labels are determined by the knowledge engineer. 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. from text unit descriptors) that are prevailing in each "acceptable" cluster. 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. Additionally, XML tags are enhanced by attributes reflecting previously extracted named entities and their values. Figure 4 contains an excerpt of the flat, unstructured and thus preliminary XML DTD that was automatically derived from XML tags in the case study. It coarsely describes the semantic structure of the resulting XML collection. Currently, named entities that serve as additional attributes of XML tags are not semantically labeled by the DIAsDEM Workbench.

In order to evaluate the quality of our approach in absence of pre-tagged documents, we drew a random sample containing 5% out of 10,785 text units and asked a domain specialist to verify the annotations of these text units with respect to the following error types:

- *Error type I:* A text unit is annotated with a wrong XML tag, i.e. the tag does not properly reflect the content of the text unit.
- *Error type II:* 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.

Within the sample, error type I (error type II) occurred in 0.4% (3.6%) of text units. Hence, tagged text units are most likely to be correctly processed. The percentage of error type II text units is higher, indicating that some text units were not placed in the cluster they semantically belong to. With 0.95 confidence, the overall error rate in the entire dataset is in the interval [2.6%, 5.9%] which is a promising result. Within this case study, the derived text unit clusterer for the Potsdam archive was also applied to semantically annotate Commercial Register entries from a different district court as well. A detailed description of the entire case study that includes the second, batch-oriented application phase of the DIAsDEM framework can be found in [32].

4.2 Planned Extensions of the DIAsDEM Framework

The development of techniques to structure the preliminary XML DTD as well as the evaluation of the extended DIAsDEM framework within sophisticated, commercially relevant case studies are the main research objectives that remain to be solved within the next year.

As summarized in the previous subsection, the DIAsDEM Workbench currently derives a flat, unstructured DTD that semantically describes an archive of XML documents. However, this rather preliminary DTD is not sufficient for subsequent usage in knowledge management or information integration efforts. Hence, structuring the derived preliminary, flat XML DTD constitutes the main current work. For that purpose, we are going to employ association rule discovery and sequence mining techniques. Since all tags are discovered by data mining techniques, we have introduced the notion of a "probabilistic DTD" to cater for inevitable tagging errors. A probabilistic DTD can be established by (i) computing statistical properties of XML tags and (ii) deriving the most likely ordering of DTD elements. We also intend to use NLP techniques and n-gram clustering as well as hierarchical clustering algorithms to discover nested DTD elements. Additionally, existing clustering algorithms and similarity measures must be evaluated with respect to the objectives of the DIAsDEM framework.

The objectives of establishing a probabilistic DTD are the specification of the most appropriate ordering of tags, the identification of correlated or mutually exclusive tags and the adornment of each tag and each correlation among tags with statistical properties. For example, the following statistical properties of DTD elements have been defined [30]:

- *Accuracy* is the probability that an XML tag correctly reflects the content of its text unit. This notion refers refers to the previously defined error type I. It is determined by domain experts who are supported by the DIAsDEM Workbench. The *Accuracy* value affects the DTD as a whole instead of being peculiar to individual tags.
- The *TagSupport* of XML tag x is defined by the ratio of XML documents that contain x to the total number of documents in the archive. This statistical property can be computed by simple frequency counts and it is peculiar to tag x. *TagSupport* is an indicator of whether an XML tag might be considered as mandatory in the DTD.
- The *AssociationConfidence* of XML tag x given the set of tags $y_1, \ldots, y_n$ is defined by the ratio of XML documents that contain the tags $y_1, \ldots, y_n$ and x to the documents containing $y_1, \ldots, y_n$. This statistical property can be computed by association rule discovery. *AssociationConfidence* is used to identify correlated tags within the archive.
- The *LocationConfidence* of tag x given the sequence of adjacent tags $y_1 \cdot y_2 \cdot \ldots \cdot y_n$ is defined by the ratio of XML documents that contain the sequence $y_1 \cdot y_2 \cdot \ldots \cdot y_n \cdot x$ to the documents containing $y_1 \cdot y_2 \cdot \ldots \cdot y_n$. This statistical property takes ordering of tags into account and can thus be discovered by sequence mining.

We have proposed to employ a directed graph to represent a probabilistic DTD. Its nodes are individual XML tags, sequences of adjacent XML tags or sets of co-occurring XML tags. Each node is adorned with statistical properties pertinent to a tag, a set or a sequence of tags. Each edge represents a relationship between XML tags, sets or sequences of XML tags. Similarly to nodes, an edge is adorned with the statistics of the order-insensitive or order-sensitive association it represents. Given appropriate thresholds, a probabilistic DTD reflects the relationships usually present in documents rather than rare ones.

Commercial Register entries are composed of rather regular and antiquated German language, which might contribute to the low overall error rate in the first case study. Therefore, this doctoral work will also include case studies from different application domains characterized by a greater linguistic diversity. For example, ad hoc news are issued by publicly quoted companies and contain 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. In the context of competition intelligence, a third case study is aimed at semantically tagging press releases, Web documents and other publicly available texts in cooperation with an pharmaceutical company. Particularly in Europe, the semantic tagging of multilingual texts is another challenge. Provided with a multilingual thesaurus and a language identifier for each document, the DIAsDEM framework should be general enough to be applied to this type of archives.

Ultimately, this doctoral work supports the project DIAsDEM by enabling the integration of archives described by probabilistic XML DTDs with related data sources [31]. In DIAsDEM, all three semantically tagged text archives will be integrated with related relational data (e.g., entries in yellow pages) into a homogeneous and commercially highly important information system. An appropriate XML query language along with a Web-based user interface will support both content- and structure-based queries that exploit the derived DTDs.

5 Conclusion

In the information age, public and private information systems frequently contain unstructured text documents of great potential value. Within this doctoral research, a new framework is being designed, implemented and evaluated that semi-automatically explicates semantic knowledge about large and domain-specific text archives to facilitate efficient text usage. The framework utilizes a knowledge discovery methodology to derive an appropriately structured, semantic XML DTD and to semantically annotate text documents with XML tags.

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] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, New York, 1999.

[3] I. Bruder, A. Düsterhöft, M. Becker, J. Bedersdorfer, and G. Neumann. GETESS: Constructing a linguistic search index for an Internet search engine. In M. Bouzeghoub, Z. Kedad, and E. Metais, editors, *Natural Language Processing and Information Systems*, number 1959 in Lecture Notes in Computer Science, pages 227–238. Springer-Verlag, 2001.

[4] 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.

[5] J. Dörre, P. Gerstl, and R. Seiffert. Text mining: Finding nuggets in mountains of textual data. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 398–401, San Diego, CA, USA, August 1999.

[6] A. Edmunds and A. Morris. The problem of information overload in business organisations: A review of the literature. *International Journal of Information Management*, 20(1):17–28, February 2000.

[7] M. Erdmann, A. Maedche, H.-P. Schnurr, and S. Staab. From manual to semi-automatic semantic annotation: About ontology-based text annotation tools. *ETAI Journal - Section on Semantic Web*, 6, 2001. To appear.

[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, 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.

[10] N. Fuhr and K. Großjohann. XIRQL – an extension of XQL for information retrieval. In *Proceedings ACM SIGIR 2000 Workshop on XML and Information Retrieval*, Athens, Greece, July 2000.

[11] H. Graubitz, 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*, pages 171–178, San Jose, CA, USA, November/December 2001.

[12] 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.

[13] P. A. Laur, F. Masseglia, and P. Poncelet. Schema mining: Finding regularity among semistructured data. In D. A. Zighed, J. Komorowski, and J. Żytkow, editors, *Principles of Data Mining and Knowledge Discovery: 4th European Conference, PKDD 2000*, volume 1910 of *Lecture Notes in Artificial Intelligence*, pages 498–503, Lyon, France, September 2000. Springer, Berlin, Heidelberg.

[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] P. Lyman and H. R. Varian. How much information. Retrieved from http://www.sims.berkeley.edu/how-much-info on 2002-05-01, 2000.

[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] G. W. Moore and J. J. Berman. Medical data mining and knowledge discovery. In *Anatomic Pathology Data Mining*, volume 60 of *Studies in Fuzziness and Soft Computing*, pages 72–117, Heidelberg, New York, 2001. Physica-Verlag.

[19] 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.

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

[21] J. F. J. Nunamaker, M. Chen, and T. D. M. Purdin. Systems development in information systems research. *Journal of Management Information Systems*, 7(3):89–106, 1990.

[22] L. Palopoli, G. Terracina, and D. Ursino. A graph-based approach for extracting terminological properties of elements of XML documents. In *Proceedings of the 17th International Conference on Data Engineering*, pages 330–337, Heidelberg, Germany, April 2001.

[23] J. Piskorski and G. Neumann. An intelligent text extraction and navigation system. In *Proceedings of the 6th International Conference on Computer-Assisted Information Retrieval*, Paris, France, 2000.

[24] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[25] 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.

[26] 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.

[27] D. Sullivan. *Document Warehousing and Text Mining*. John Wiley & Sons, New York, Chichester, Weinheim, 2001.

[28] 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.

[29] 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.

[30] K. Winkler and M. Spiliopoulou. Extraction of semantic XML DTDs from texts using data mining techniques. In *Proceedings of the K-CAP 2001 Workshop on Knowledge Markup and Semantic Annotation*, pages 59–68, Victoria, BC, Canada, October 2001.

[31] K. Winkler and M. Spiliopoulou. Integrating data and probabilistically structured text documents. In *Proceedings des 5. Workshops "Förderierte Datenbanken" und GI Arbeitstreffen "Konzepte des Data Warehousing" (FDBS 2001)*, pages 16–29, Berlin, Germany, October 2001.

[32] K. Winkler and M. Spiliopoulou. Semi-automated XML tagging of public text archives: A case study. In *Proceedings of EuroWeb 2001 "The Web in Public Administration"*, pages 271–285, Pisa, Italy, December 2001.