Employing Text Mining for Semantic Tagging in DIAsDEM

Karsten Winkler[1] and Myra Spiliopoulou

Myra Spiliopoulou studied Mathematics, received her Ph.D. in Computer Science (1992, University of Athens) and obtained her Habilitation in Information Systems (2000, Humboldt University Berlin). Since April 2001, she is Professor of E-Business in the Leipzig Graduate School of Management. Her research interests include web mining, text mining and knowledge management in information systems. Karsten Winkler studied Business Administration (1999, Humboldt University Berlin). Currently, he is a research assistant in Myra Spiliopoulou's department and is interested in text mining, information fusion and web mining.

Abstract

Both public and private organizations have been accumulating large volumes of electronically available text documents for the past years. However, to turn text archives into profitable sources of knowledge, they should be transformed into an integrated and efficiently queryable information system. To attain this objective, the project DIAsDEM employs data mining techniques to derive a semantic XML DTD for a text archive and to semantically annotate its documents. In this article, we briefly describe the DIAsDEM framework for semantic tagging and its application in a case study.

1. Introduction

Most organizations are not only "drowning" in data, they are also "struggling" to cope with huge amounts of text documents. Tan points out that up to 80% of a company's information is stored in unstructured textual documents [5]. Hence, capturing interesting and actionable knowledge from textual databases is a major challenge for the data mining community. Creating semantic markup is one form of providing explicit knowledge about text archives to facilitate searching and browsing or to enable information integration with related data sources. Unfortunately, most users are not willing to manually create meta-data due to the efforts and costs involved [1]. Thus, text mining techniques are required that (semi-) automatically create semantic markup.

In the research project DIAsDEM, we are exploring KDD techniques for the integration of texts and related structured data. The project whose German acronym stands for "Data Integration of Legacy Systems and Semi-Structured Documents by Means of Data Mining Techniques" is funded by the German Research Society (DFG). The research groups of Prof. Dr. Stefan Conrad at the Ludwig Maximilian University Munich (Computer Science Department) and Prof. Dr. Myra Spiliopoulou at the Leipzig Graduate School of Management (Department of E-Business) are cooperating on this project. Its main objective is the incorporation of legacy 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 should be identified in unstructured or semi-structured documents. Secondly, dependencies among attributes of different legacy data sources must be detected. Afterwards, the results can be applied to integrate related data from various heterogeneous sources.

---

[1] The work of this author is funded by the German Research Society (DFG grant no. SP 572/4-1).

In the next section, we briefly introduce our framework to solve the first DIAsDEM research issue. Our KDD approach aims at deriving a preliminary flat XML DTD serving as a quasi-schema for the document archive and at enabling the provision of database-like querying services on textual data. The reader might refer to [3,2] for a complete description of the DIAsDEM framework for semantic tagging of domain-specific texts as well as a thorough discussion of related work. The case study is described in detail in [8]. In the last section, we conclude and give directions for future research in the project.

2. The DIAsDEM Framework

In our project, 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. 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 attributes describing extracted named entities (e.g., persons). The XML document 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 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 business information, their textual content can only be searched using full-text queries at the moment. Hence, semantically semi-structuring these textual archives provides a basis for information integration and value-adding services.

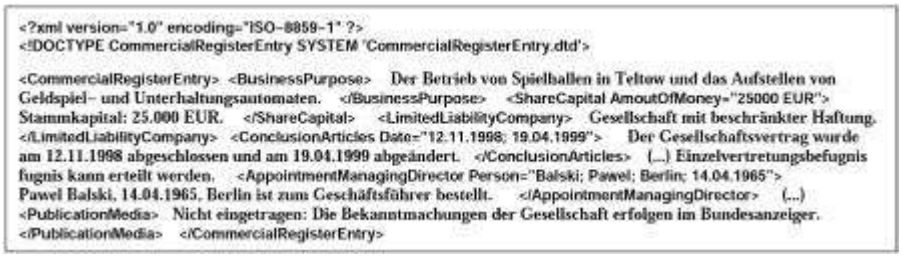

Figure 1: XML document containing an annotated Commercial Register entry

Our 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. Semantic tagging is a two-phase process in DIAsDEM. We have designed a knowledge discovery in textual databases (KDT) process that constitutes the first phase to discover 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 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.

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 published on electronic marketplaces.

The iterative and interactive KDT process that constitutes the first phase of the DIAsDEM framework is depicted in Figure 2. 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 naming decisions.

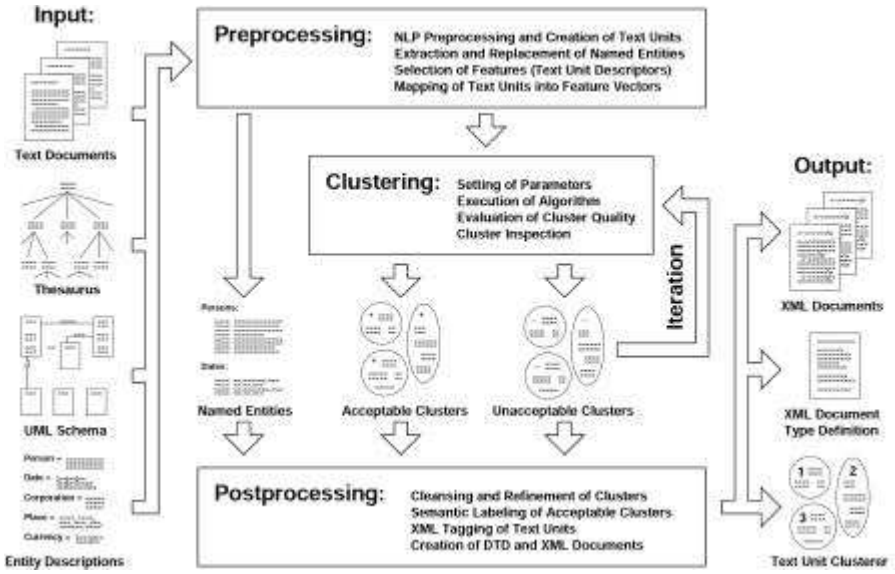

Figure 2: Iterative and interactive KDT process of the DIAsDEM framework

Besides the initial text documents to be tagged, the following domain knowledge constitutes input to our KDT 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 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 preprocessing phase: After setting the level of granularity by determining the size of text units, the Java- and Perl-based DIAsDEM Workbench performs basic NLP preprocessing such as tokenization, normalization and word stemming using TreeTagger [4]. 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 thesaurus and the UML schema. Text unit descriptors are currently chosen by the knowledge engineer because they must reflect important concepts of the application domain. All text units are mapped onto Boolean and subsequently TF-IDF weighted vectors of this feature space. Additionally, named entities of interest are extracted from text units by a separate module of the DIAsDEM Workbench. In our case study, we created a small thesaurus and selected 70 relevant descriptors and 109 non-descriptors pointing to descriptors.

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, whereas 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 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" cluster 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. 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-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. 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 the derived XML tags. Additionally, XML tags are enhanced by attributes reflecting previously extracted named entities and their values. Figure 3 contains an excerpt of the flat, unstructured XML DTD that was automatically derived from XML tags in the case study. It coarsely describes the semantic structure of the resulting XML collection.

```
<?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 3: Preliminary flat, unstructured XML DTD of Commercial Register entries

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 their annotations with respect to the following error types: Firstly, a text unit is annotated with a wrong XML tag, i.e. its tag does not properly reflect the content of the text unit (error type I). 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 (error type II). 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.

3. Conclusion

Archives of unstructured text documents often contain information of great potential value. Transforming large volumes of textual data into commercially useful sources of knowledge is currently a significant challenge. Within the scope of DIAsDEM, we proposed, implemented and evaluated an extensive framework for semantic tagging of large, domain-specific and rather homogeneous text collections. Applying this framework to an archive is a prerequisite for subsequent integration with related unstructured, semi-structured or structured data sources in a unifying information system. In the sense of information fusion, DIAsDEM ultimately aims at creating a single, transparent and value-adding information system that can be appropriately queried.

To attain this objective, many challenging research issues remain open. Our current and future work includes the transformation of a derived flat, unstructured and rather preliminary XML DTD into a

structured DTD [7]. Given that all DTD elements are derived by data mining techniques, tagging errors are very likely to occur. Thus, we introduce the notion of a probabilistic DTD describing the most likely ordering of DTD elements and containing statistical properties of its elements. Secondly, we are preparing the integration of archives described by probabilistic DTDs with related data sources [6]. Additionally, we will evaluate clustering algorithms, similarity metrics and XML query languages with respect to our objectives.

Commercial Register entries are composed of rather formal and rigid language. Hence, they represent an ideal domain for validating our framework. However, we are currently investigating the semantic tagging of linguistically more diverse ad hoc news that are issued by publicly quoted companies. They contain important information that might influence share prices. As the first results are very promising, we are planning to integrate Commercial Register entries, ad hoc news and relational data contained in publicly available "Yellow Pages" in an information system supporting business decisions.

References

[1] 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.

[2] 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, San Jose, CA, USA, November/December 2001. To appear.

[3] 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, Gommern, Germany, May 2001.

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

[5] 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.

[6] Karsten Winkler and Myra 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.

[7] Karsten Winkler and Myra 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.

[8] Karsten Winkler and Myra Spiliopoulou. Semi-automated XML tagging of public text archives: A case study. In *Proceedings of EuroWeb 2001 "The Web in Public Administration"*, Venice, Italy, December 2001. To appear.