VALIDATION OF MIXED-STRUCTURED DATA USING PATTERN MINING AND INFORMATION EXTRACTION

Martin Atzmueller

University of Kassel
Knowledge and Data Engineering
Kassel, Germany
atzmueller@cs.uni-kassel.de

Stephanie Beer

University-Hospital of Würzburg
Gastroontologics Research Group
Würzburg, Germany
beer_s@klinik.uni-wuerzburg.de

ABSTRACT

For large-scale data mining utilizing data from ubiquitous and mixed-structured data sources, the appropriate extraction and integration into a comprehensive data-warehouse is of prime importance. Then, appropriate methods for validation and potential refinement are essential. This paper presents an approach applying data mining and information extraction methods for data validation: We apply subgroup discovery and (rule-based) information extraction for data integration and validation. The methods are integrated into an incremental process for continuous validation options. The results of a medical application demonstrate that subgroup discovery and the applied information extraction methods are well suited for mining, extracting and validating clinically relevant knowledge.

1. INTRODUCTION

Whenever data is continously collected, for example, using intelligent documentation systems [1], data mining and data analysis provide a broad range of options for scientific purposes. The mining and analysis step is often implemented using a data-warehouse [2, 3, 4]. For the data preprocessing and integration of several heterogenous sources, there exist standardized extract-transform-load (ETL) procedures, that need to incorporate suitable data schemas, and integration rules. Additionally, for unstructured or semi-structured textual data sources, the integration requires effective information extraction methods. For clinical discharge letters, for example, the structure of the letter is usually non-standardized, and thus dependent on different writing styles of different authors.

However, a prerequisite of data mining is the validation and the quality assurance of the integrated data. Especially concerning unreliable extraction and integration methods, the quality of the obtained data can vary significantly. If the data has been successfully validated, then the trust in the data mining results and their acceptance can be increased.

In this paper, we propose an approach for the validation of mixed-structured data using data mining and information extraction and propose appropriate refinement options. We focus on a data mining technique for mining *local patterns*, i.e., subgroup discovery, e.g., [5, 6, 7] that are especially suitable for the task: Local patterns consider local regularities (and irregularities) of the data and are therefore useful for spotting non-expected, contradicting, and otherwise unusual patterns potentially indicating problems and errors in the data.

Concerning the information extraction techniques, we consider popular methods implemented in the UIMA [8] and ClearTK [9] framework, and especially focus on the TEXTMARKER system, e.g., [10, 11] for rule-based information extraction. Rules are especially suitable for the proposed information extraction task since they allow a concise and declarative formalization of the relevant domain knowledge that is especially easy to acquire, to comprehend and to maintain. Furthermore, in the case of errors, the cause can easily be identified by tracing the application of the individual rules.

The combined approach enables data mining from heterogenous sources. The user can specify simple rules that consider features of the text, e.g., structural or syntactic features of the textual content. We focus on an incremental level-wise approach, such that both methods can complement each other in the validation and refinement setting. Furthermore, validation knowledge can be formalized in a knowledge base, for assessing known and expected relations in the data.

The approach has been implemented in a clinical application for mining data from clinical information systems, documentation systems, and clinical discharge letters. This application scenario concerns the data integration from heterogenous databases and the information extraction from textual documents. The experiences and results so far demonstrate the flexibility and effectiveness of the presented approach that make the data mining and information extraction methods suitable components in the mining, validation and refinement process.

2. BACKGROUND

In the following, we shortly summarize the methods for data mining and information extraction, subgroup discovery, and rule-based information extraction using TEXTMARKER.

2.1. Subgroup Discovery

Subgroup discovery is a flexible data mining method for discovering local patterns that can be utilized for global modeling in the context of exploratory data analysis, description, characterization and classification.

Subgroup discovery is applied for identifying relations between a (dependent) target concept and a set of explaining (independent) variables. Then, the goal is to describe subsets of the data, that have the most unusual characteristics with respect to the concept of interest given by the target variable [6]. For example, the risk of coronary heart disease (target variable) is significantly higher in the subgroup of smokers with a positive family history than in the general population.

In the context of the proposed validation approach, we consider certain gold-standard concepts as targets, as well as target concepts that are true, if and only if equivalent concepts from two different sources match. Then, we can identify combinations of factors that cause a mismatch between the concepts. These combinations can then indicate candidates for refinement.

2.2. Rule-based Information Extraction

Information extractions aims at extracting a set of *concepts*, *entities* and *relations* from a set of documents. TEXTMARKER [10, 11] is a robust system for rule-based information extraction. It can be applied very intuitively, since the used rules are especially easy to acquire and to comprehend. Using the extracted information, data records can be easily created in a post-processing step. Humans often apply a strategy according to a *highlighter metaphor* during 'manual' information extraction: First, top-level text blocks are considered and classified according to their content by coloring them with different highlighters. The contained elements of the annotated texts segments are then considered further. The TEXTMARKER [10, 11] system tries to imitate this manual extraction method by formalizing the appropriate actions using *matching rules*: The rules mark sequences of words, extract text segments or modify the input document depending on textual features.

TextMarker aims at supporting the knowledge engineer in the rapid prototyping of information extraction applications. The default input for the system is semi-structured text, but it can also process structured or free text. Technically, HTML is often the input format, since most word processing documents can be obtained in HTML format, or converted appropriately.

3. THE MINING AND VALIDATION PROCESS

Figure 1 depicts the process of validation and refinement of mixed-structured data using pattern mining and information extraction methods. The input of the process is given by data from heterogenous data sources, and by textual documents. The former are processed by appropriate data integration methods adapted to the different sources. The latter are handled by information extraction techniques, e.g., rule-based methods that utilize appropriate extraction rules for the extraction of concepts and relations from the documents. In general, a variety of methods can be applied.

The process supports arbitrary information extraction methods, e.g., automatic techniques like support-vector machines or conditional random fields as implemented in the ClearTK [9] toolkit for statistical natural language processing. However, the refinement capabilies vary for the different extraction approaches: While black-box methods like support vector machines or conditional random fields only allow an indirect refinement and adaptation of the model, i.e., based on adapting the input data and/or the method parameters for constructing the model, a white-box approach implemented using rules provides for a direct modification of its model, namely the provided rules. Therefore, we especially focus on rule-based methods due to their rich refinement capabilities.

After the integration and extraction of the data, the result is provided to the pattern mining system which obtains a set of validation patterns as output. This set is then checked both for internal consistency and compared to formalized background knowledge. In the case of discrepancies and errors, refinement are proposed for the data integration and/or the information extraction steps. After the rules have been refined, the process iterates with the updated schemas and models.

In the following we discuss exemplary results obtained from a medical project. We applied data collected by the SONOCONSULT system, a multifunctional knowledge system for sonography, which has been in routine use since 2002 documenting more than 12000 patients in two clinics. The system covers the entire field of abdominal ultrasound (liver, portal tract, gallbladder, spleen, kidneys, adrenal glands, pancreas, intestine, lymph nodes, abdominal aorta, cava inferior, prostate, and urinary bladder). The data was integrated with the SAP-based i.s.h.med system, and the information extraction techniques were applied for textual discharge letters from the respective patients; SONOCONSULT was used for documentation. By integrating different data sources into the warehouse it is possible to measure the conformity of sonographic results with other methods or inputs. In our evaluations, we applied computer-tomography diagnoses and additional billing diagnoses (from the hospital information system) as a gold-standard.

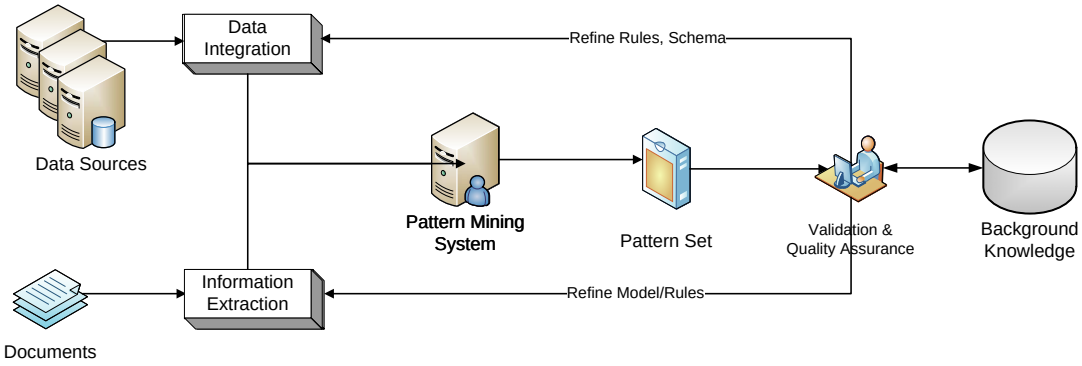

Fig. 1. Process Model: Validation of Mixed-Structured Data using Pattern Mining and Information Extraction

Table 1 shows the correlation of SONOCONSULT based diagnosis with CT/MR, diagnoses listed in the discharge letter and diagnoses contained in the hospital information system for a selection of cases from a certain examiner. It was quite interesting that the conformity between SONOCONSULT based diagnoses with the diagnoses contained in the hospital information system was relatively low. Evaluating this issue it was obvious that various diagnosis were not listed in the hospital information system because they were not revenue enhancing and not relevant for all clinical situations. Therefore, we looked at the accordance with the discharge letters which were found to be highly concordant at least for the diagnosis of liver metastasis. Liver cirrhosis is more awkward to detect using ultrasound and has to be in a more advanced stage. Therefore, some of the discharge diagnoses "liver cirrhosis" were only detected using histology or other methods.

In some cases, there are discrepancies with respect to the formalized background knowledge that still persist after refinement of the rules and checking the data sources. In such cases, explanation-aware mining and analysis components provide appropriate solutions for resolving conflicts and inconsistencies. By supporting the user with appropriate justifications and explanations, misleading patterns can be identified, and the background knowledge can be adapted. The decision whether the background knowledge needs to be adapted is performed by the domain specialist. As we have described in [12] there are several continuous explanation dimensions in the context of data mining and analysis, that can be utilized for improving the explanation capabilities. In the medical domain, for example, patterns are usually first assessed on the abstract level, before they are checked and verified on concrete patient records, i.e., on a very detailed level of abstraction. Then, discrepancies are modeled in the background knowledge, for example, certain exception conditions for certain subgroups of patients.

The validation phase is performed on several levels: On the first level, we can use a (partial) gold-standard both for checking the data integration and information extraction tasks. We only require a partial gold-standard, i.e., a sample of the correct relations, because we need to test the functional requirements of the data integration and extraction phases. On the next level, we can incrementally validate the integrated data using the extracted information, or vice versa, using the mined patterns. In the case of discrepancies, we can rely on the partial gold-standard data for verification, or we can identify potential causes and verify these on concrete cases. Therefore, the final decision for the refinements relies on the user, which reviews all proposed refinements in a semi-automatic approach.

For the refinement steps, we can either extend the (partial) gold-standard, or we perform a boot-strapping approach, using a small gold-standard sample of target concepts for validation, e.g., for validating and refining the information extraction approach, which is in turn used for the validation of the data sources. In the next step, the validation targets can be extended and the process for refinement is applied inversely. The boot-strapping approach for validation and refinement is thus similar to the idea of co-training, e.g., [13] in machine learning that also starts with a small labeled (correct) dataset and iteratively adapts the models using another co-trained dataset.

4. CONCLUSIONS

This paper presented an approach for the validation of mixed-structured data using information extraction and pattern mining methods. In an incremental approach, data can both be validated and refined with an increasing level of accuracy. The presented approach has been successfully implemented in a medical project targeted at integrating data from clinical information systems, documentation systems, and textual discharge letters.

The experiences and results so far demonstrate the flexibility and effectiveness of the pattern mining and information extraction methods for the presented validation and refinement approach.

Total Case Number	SONO CONSULT Diagnoses	SAP Diagnoses	% Conformity with SONO CONSULT	CT/MR Diagnoses	% Conformity with SONO CONSULT	Discharge Letter Diagnoses	% Conformity with SONO CONSULT
Liver cirrhosis							
16	12	6	20	1	33	9	50
Liver metastasis							
28	16	11	65	15	87	17	94

Table 1. Exemplary study for a selection of cases concerning liver examinations performed by a certain examiner: Conformity of system diagnoses with various sources of diagnosis input. The columns indicate the degree of correlation of the different sources with SONOCONSULT diagnoses measured by the number of covered cases.

5. REFERENCES

[1] Frank Puppe, Martin Atzmueller, Georg Buscher, Matthias Huettig, Hardi Lührs, and Hans-Peter Buscher, "Application and Evaluation of a Medical Knowledge-System in Sonography (SonoConsult)," in *Proc. 18th Europ. Conf. on Artificial Intelligence (ECAI 2008)*, 2008, pp. 683–687.

[2] Jonathan C. Prather, David F. Lobach, Linda K. Goodwin, Joseph W. Hales, Marvin L. Hage, and W. Edward Hammond, "Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse," in *Proc. AMIA Annual Fall Symposium (AIMA-1997)*, 1997, pp. 101–105.

[3] Rüdiger Wirth and Jochen Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining," in *Proc. 4th Intl. Conf. on the Practical Application of Knowledge Discovery and Data Mining*. 2000, pp. 29–39, Morgan Kaufmann.

[4] Martin Atzmueller, Stephanie Beer, and Frank Puppe, "A Data Warehouse-Based Approach for Quality Management, Evaluation and Analysis of Intelligent Systems using Subgroup Mining," in *Proc. 22nd International Florida Artificial Intelligence Research Society Conference (FLAIRS), accepted*. 2009, pp. 372–377, AAAI Press.

[5] Martin Atzmueller, Frank Puppe, and Hans-Peter Buscher, "Exploiting Background Knowledge for Knowledge-Intensive Subgroup Discovery," in *Proc. 19th Intl. Joint Conference on Artificial Intelligence (IJCAI-05)*, Edinburgh, Scotland, 2005, pp. 647–652.

[6] Stefan Wrobel, "An Algorithm for Multi-Relational Discovery of Subgroups," in *Proc. 1st European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD-97)*, Berlin, 1997, pp. 78–87, Springer Verlag.

[7] Willi Klösgen, "Explora: A Multipattern and Multistrategy Discovery Assistant," in *Advances in Knowledge Discovery and Data Mining*, Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padraic Smyth, and Ramasamy Uthurusamy, Eds., pp. 249–271. AAAI Press, 1996.

[8] David Ferrucci and Adam Lally, "UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment," *Nat. Lang. Eng.*, vol. 10, no. 3-4, pp. 327–348, 2004.

[9] P. V. Ogren, P. G. Wetzler, and S. Bethard, "ClearTK: A UIMA Toolkit for Statistical Natural Language Processing," in *UIMA for NLP workshop at Language Resources and Evaluation Conference (LREC)*, 2008.

[10] Martin Atzmueller, Peter Kluegl, and Frank Puppe, "Rule-Based Information Extraction for Structured Data Acquisition using TextMarker," in *Proc. of the LWA-2008, Special Track on Knowledge Discovery and Machine Learning*, 2008, pp. 1–7.

[11] Peter Kluegl, Martin Atzmueller, and Frank Puppe, "Textmarker: A tool for rule-based information extraction," in *Proc. Biennial GSCL Conference 2009, 2nd UIMA@GSCL Workshop*. 2009, pp. 233–240, Gunter Narr Verlag.

[12] Martin Atzmueller and Thomas Roth-Berghofer, "Ready for the MACE? The Mining and Analysis Continuum of Explaining Uncovered," in *AI-2010: 30th SGAI International Conference on Artificial Intelligence. Accepted.*

[13] Avrim Blum and Tom Mitchel, "Combining Labeled and Unlabeled Data with Co-Training," in *COLT: Proceedings of the Workshop on Computational Learning Theory*. 1998, pp. 92–100, Morgan Kaufmann.