Wiki-Enabled Semantic Data Mining – Task Design, Evaluation and Refinement

Martin Atzmueller, Florian Lemmerich, Jochen Reutelshoefer, and Frank Puppe

University of Würzburg,
Department of Computer Science,
Am Hubland, 97074 Würzburg, Germany
lastname@informatik.uni-wuerzburg.de

Abstract. Complementing semantic data mining systems by wikis and especially semantic wikis yield a flexible knowledge-rich method. This paper describes a system architecture of a collaborative approach for semantic data mining. The goal is to enhance the design, evaluation and refinement of data mining tasks using semantic technology. Collaborative aspects are introduced by utilizing wiki technology. We present the components and describe their interaction and application in detail.

1 Introduction

Wikis provide flexible ways for supporting the quick and simple creation, sharing and management of content. Based upon the established wiki-technology, semantic wikis (e.g., [1,2]) enhance this by providing enriched content and features. For example, flexible inline queries and according results that are generated based on these dynamically are such prominent features. While the queries and answers (results) can be flexibly handled by the system, and can usually be formalized as textual content, the wiki system also provides appropriate means for the persistent storage and management of the generated content.

Semantic data mining systems enable the inclusion of a large set of background knowledge, for example, in order to access knowledge services, for selecting the applied data mining methods, or for postprocessing the obtained data mining results. Thus, integrating wikis is a convenient option for semantic data mining systems, since the semantic core components can support the semantic mining features, while the wiki component provides for a convenient front-end and user-management, enables the persistent storage of queries and mining results, and supports their extended annotation.

This paper presents a wiki-enabled approach for collaborative semantic data mining: The semantic data mining system VIKAMINE [3] is combined with the semantic JSPWiki (`http://www.jspwiki.org`) extension KnowWE [1]. We describe the interaction and exchange of query and results data, and the integration of semantic information and knowledge.

The rest of the paper is structured as follows: Section 2 describes the basics of semantic data mining, provides an overview of the presented approach, and describes its implementation. Section 3 concludes with a summary and interesting directions for future work.

2 Method

This section briefly introduces the general semantic data mining approach. After that, we first give a general overview, before we describe the architecture of the proposed approach in detail. Finally, we discuss related work.

2.1 Semantic Data Mining

Semantic data mining can be considered as an approach utilizing formal methods and techniques in order to explicitly integrate data semantics, background knowledge, or reasoning in the mining process. The knowledge is typically represented in a knowledge repository, such as an ontology, or a knowledge base. The main aspect of semantic data mining is the explicit integration of this knowledge into the data mining and knowledge discovery process, where the algorithms for data pre-processing, mining or post-processing make use of the formalized knowledge to improve the overall process. There has been growing interest in this issue, e.g., [4–6], in various domains, especially in the medical domain [4, 7, 8].

With the advent of the semantic web and standardized knowledge representations of semantic web techniques, e.g., the web ontology language *OWL*, utilizing these knowledge representation formalisms for data mining is a promising direction for task design, evaluation and refinement, as discussed below. In the following, we outline the different aspects of semantic data mining, and discuss their implications.

The general data mining process can be structured along the CRISP-DM process model (`http://www.crisp-dm.org`) and consists of the following phases: (a) Business understanding, i.e., understanding the application domain, (b) Data Understanding, i.e., considering the (potential) objects of analysis, (c) Data Preparation, e.g., preprocessing and schema-matching of the data elements, (d) Modeling, e.g., given by concrete mining sessions, (e) Evaluation, i.e., assessment of the mined models, (f) Deployment, i.e., putting the extracted knowledge into action. The semantic data mining approach integrates ontologies in each of the six steps [7]. In the following, we provide examples for each of the phases structured along the dimensions of task design, evaluation, and refinement.

- **Task Design**:
  - In the **Business Understanding** phase ontologies help inexperienced users getting accostumed to the domain, by structuring the relations between the concepts, and explaining the concepts in terms of their properties.
  - In the **Data Understanding** phase, important data elements (contained in the ontology) need to be selected. Then, missing attributes, or redundant attributes can be added or removed from the data set. This can be accomplished by a *data-to-ontology mapping* step [5] where the data elements are mapped to concepts of the ontology, e.g., for integrating heterogenous data.
  - The **Data Preparation** phase is strongly connected to the *Modeling* phase. Depending on the latter, for example constraints on attributes or values can be derived. This concerns constraints on the relations between the attributes, as described in [4], for example, grouping constraints or exclusion constraints for

  certain attribute groups that should not be considered. A further possible inclusion of the ontology is given by a more abstract task composition phase, for which the modeling phase can be hierarchically decomposed along the generalization/specialization hierarchies modeled in the ontology. Then, more concise results can potentially be obtained on lower levels, but for efficiency reasons higher levels can be considered first and be used for filtering interesting hypotheses in an earlier stage, cf., [5].

- **Task Evaluation**:
  - During the **Evaluation** phase (of CRISP-DM), the discovered patterns can be interpreted and explained in a structured way using the concepts and/or contained patterns. Various post-processing options are available at this point, cf., [5]. Specifically, due to the data-to-ontology mapping, the discovered patterns can be matched to semantic relations or more complex relations between these. Additionally, such knowledge provides a potential (explaining) context for the discovered patterns. Furthermore, prior knowledge can be compared to the patterns, e.g., for confirming known relations, identifying new knowledge, and/or detecting exceptions and conflicts with formalized expectations. Concerning possible explanations, causal relations can often help in this respect, for validating and confirming discovered patterns, or for their analysis.
  - The **Deployment** phase concerns the integration of the discovered models into the business setting. It is easy to see, that for distributed processing and storage (e.g., on the semantic web) a shared ontology is inevitable. This is especially relevant for deploying results as *semantic analytic reports* (an extension of *analytic reports* [5]), described below. In a late evaluation step, the models/patterns can be tested during their practical application. In that case, the persistent sessions stored in the wiki provide direct access in a collaborative manner.
- **Task Refinement**: The task refinement step is activated after the evaluation step has been performed. It is accomplished either manually using the wiki system – by modifying the textual task description, or by applying formalized knowledge with respect to the applied data mining method. Then, parameters and/or the method itself can be adapted. Refinement is performed according to the results of the *evaluation* phase, so both steps are tightly coupled. Due to the application of the wiki, different persons can collaborate in separate sessions, such that previous results can be included in the refinement of other (related) sessions. Furthermore, previous experiences can be documented using the wiki, for example, explanations/comments by previous users. Furthermore, special refinement and/or evaluation knowledge can be formalized for further improving the respective steps.

2.2 General Overview

As discussed in the last section *semantic data mining* is concerned with the utilization of ontological knowledge and semantic annotations to be used throughout the data mining and knowledge discovery process, similar to *ontology-enhanced* [5] data mining. However, further semantic features are enabled by including a *semantic core* component, e.g., a RDF-Store: Using that, results can be incrementally formalized and provided to the store, while subsequent mining and semantic queries can make use of the collected knowledge. The data mining query, results, and additional knowledge can then be transparently integrated into a *semantic analytic report*: The idea of such reports is based on *analytical reports* [5] that are simple text documents containing the mining results with additional text (which is created by humans). In the semantic setting, we can automatically transform the mining results into a format suitable for the report. Additionally, the content can be enriched using semantic annotations and links between the reports (and background information). The wiki also provides for flexible versioning which is especially useful in a collaborative setting.

The sketched scenario is especially suitable for inexperienced users that are mainly interested in reporting features of a data mining system. Such reports provide high-level access to pre-specified queries that can be evaluated routinely. However, using the wiki query mechanism, such queries can also be formalized in an ad-hoc fashion. Further more detailed reports, analyses and mining sessions can then be implemented using more advanced data mining tools, e.g., by applying the VIKAMINE [3] system.

On the application side, specialized sessions with domain experts, e.g., medical doctors, and data mining engineers can be easily implemented using the collaborative tool. In this context, the proposed approach provides, for example, flexible query formalization, versioning, a history of queries and results, and the potential for knowledge and experience management since the obtained semantic analytical reports can be commented on, and can be linked to other (similar) documents. Further sessions can thus easily build on results of previous sessions, with the same or new participants. For experience management, the wiki can also be combined with a tagging system, e.g., [9].

2.3 Architectural Overview

Due to the limited space, we only provide a brief summary of the architecture of the proposed approach. A more comprehensive discussion and overview is given in [10]. The architecture consists of two core components: The basic wiki system (provided by JSPWiki (`http:/www.jspwiki.org`) is extended by the semantic wiki extension KnowWE [1]. The wiki component provides basic features like editing, versioning, user management, access management and attachment management. Additionally, it directly supports the collaborative aspects of the sketched semantic mining approach. KnowWE itself is designed as a highly extensible minimal core providing basic semantic wiki features like formalization and reasoning. Therefore, for communication with the mining component we designed the connector plugin *KnowWE-RIP* (REST [11] Interface Plugin) that facilitates the connection to the mining web-service. The semantic core component for storage and reasoning is given by a combination of the *Sesame* (`http://www.openrdf.org`) framework and OWLIM. Sesame is a java-based framework with support for storing and analyzing RDF data. OWLIM is a semantic repository with reasoning capabilities that is packaged as a storage and inference layer for sesame. As such, KnowWE integrates a semantic component and contains a connector to the Sesame/OWLIM components for providing the semantic functionality.

We utilize the VIKAMINE [3] system (`http://www.vikamine.org`) for data mining. VIKAMINE features a web-service that can be queried using XML based on a specialized query language. The result (i.e., the answer) is also formulated as XML and can thus transparently be integrated with the wiki.

The semantic mining process is initiated by the user, that is, by formulating a query to the wiki system. Similar to other wiki-systems, the query is provided in the form of an *inline-query* (e.g., [2]): The query is directly entered in textual form. Whenever the wiki page is stored and/or reloaded with a new or modified query the result is requested. In addition, we provide 'extended' inline queries, such both the query and the result (i.e., the 'answer') can also be shown as required. Technically, the query is first transformed to an XML-representation (VPDL, the VIKAMINE *Pattern Description Language*, and then forwarded to the mining engine that produces an result in XML/VPDL format. Finally, this result is re-transformed into human-readable textual form to be displayed by the wiki. However, internally the 'raw' result can be retained by the versioning system of the wiki, such that always the latest result is available and can be cached for efficiency. Therefore, changes, for example, due to an updated dataset, can be easily extracted. The general architecture is shown in Figure 1. The seamless integration of the result presentation enables (inexperienced) users to quickly evaluate the obtained results by themselves and according to the formalized ontological knowledge.

2.4 Related Work and Discussion

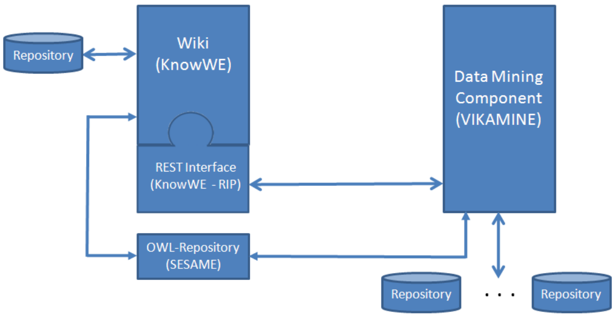

Fig. 1. Semantic Data Mining Architecture

Using ontologies for enhancing data mining has been discussed, e.g., by Svatek et al. [5] and by Antunes [6] in the context of mining association rules. Furthermore, Cespivova et al. [7] and Kuo et al. [8] describe applications in the medical domain. While the application of ontologies is also a focus of the presented approach, the proposed method aims at a more comprehensive integration of semantic information and knowledge. In contrast to the existing approaches, the proposed approach considers a comprehensive *two-way* integration of semantic and data mining methods for semantic data mining, with feedback in both directions. In this way, prior knowledge can be transparently integrated. Using the wiki-support of the presented approach collaborative sessions can be implemented. Furthermore, semantic annotations using the wiki, linking unstructured, semi-structured and structured information is another novel issue with respect to the presented approach. Semantic analytical reports can include semantic annotations at the document level, global tagging, and associated query – data mining results that are stored in the semantic store and thus provide powerful options for knowledge-rich applications.

3 Conclusions

In this paper, we have presented an approach for collaborative semantic data mining. We discussed the considerations of task design, evaluation and refinement in the context of semantic data mining. Additionally, we have presented an overview on the approach and have described its architectural considerations in detail, utilizing the VIKAMINE system for semantic data mining with a connector to a wiki system.

For future work, we aim to extend the mining approach towards text mining and information extraction, e.g., [12]. This opens up further potential for incremental knowledge refinement, discovery and semantic annotation.

Acknowledgements

This work has been partially supported by the German Research Council (DFG) under grant Pu 129/8-2.

References

1. Reutelshoefer, J., Haupt, F., Lemmerich, F., Baumeister, J.: An Extensible Semantic Wiki Architecture. In: Proc. 4th Workshop on Semantic Wikis - The Semantic Wiki Web. (2009)
2. Krötzsch, M., Vrandecic, D., Völkel, M., Haller, H., Studer, R.: Semantic Wikipedia. Web Semantics: Science, Services and Agents on the World Wide Web **5**(4) (2007) 251 – 261
3. Atzmueller, M., Puppe, F.: Semi-Automatic Visual Subgroup Mining using VIKAMINE. Journal of Universal Computer Science (JUCS), Special Issue on Visual Data Mining **11**(11) (2005) 1752–1765
4. Atzmueller, M., Seipel, D.: Using Declarative Specifications of Domain Knowledge for Descriptive Data Mining. In: Proc. 18th International Conference on Applications of Declarative Programming and Knowledge Management, Berlin, Springer Verlag (2008)
5. Svátek, V., Rauch, J., Ralbovský, M.: Ontology-Enhanced Association Mining. In: Semantics, Web and Mining. Volume 4289 of LNCS. (2005) 163–179
6. Antunes, C.: Onto4AR: A Framework for Mining Association Rules. In: International Workshop on Constraint-Based Mining and Learning (CMILE 2007), Warsaw, Poland (2007)
7. Cespivova, H., Rauch, J., Svatek, V., Kejkula, M.: Roles of Medical Ontology in Association Mining CRISP-DM Cycle. In: Proc. ECML/PKDD 2004 Workshop on Knowledge Discovery and Ontologies, Pisa, Italy (2004)
8. Kuo, Y.T., Lonie, A., Sonenberg, L., Paizis, K.: Domain Ontology Driven Data Mining: A Medical Case Study. In: DDDM '07: Proceedings of the 2007 international workshop on Domain driven data mining, New York, NY, USA, ACM (2007) 11–17
9. Atzmueller, M., Haupt, F., Puppe, F.: Knowta: Wiki-Enabled Social Tagging for Collaborative Knowledge and Experience Management. In: Proc. 2nd International Workshop on Design, Evaluation and Refinement of Intelligent Systems (DERIS). (2009)
10. Atzmueller, M., Lemmerich, F., Reutelshoefer, J., Puppe, F.: An Extensible Architecture for Wiki-enabled Semantic Data Mining. In: Technical Report. University of Wuerzburg. (2009)
11. Fielding, R.T., Taylor, R.N.: Principled Design of the Modern Web Architecture. ACM Transactions on Internet Technology **2**(2) (2002) 115–150
12. Atzmueller, M., Kluegl, P., Puppe, F.: Rule-Based Information Extraction for Structured Data Acquisition using TextMarker. In: Proc. LWA 2008 (Knowledge Discovery and Machine Learning Track), University of Wuerzburg (2008)