Integrating and Updating Domain Knowledge with Data Mining

Carsten Pohle

HHL – Leipzig Graduate School of Management
cpohle@ebusiness.hhl.de

Abstract

Most current tools for data mining lack support for intelligent analysis and filtering of mined patterns. Dividing interesting mining results from uninteresting ones still is a laborious task mainly performed by human users. We propose to employ formalized domain knowledge for assessing the interestingness of mining results. We present considerations and ideas as foundations of the design of an intelligent data mining environment.

1 Introduction

Data mining – or knowledge discovery in data bases (KDD) – is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [6]. During the past decade, this discipline has attracted the attention of many researchers in the fields of data analysis, machine learning, statistics, databases and management information systems. Today, the results of these research efforts have already made their way into practical applications for marketing, customer relationship management, fraud detection or Web usage analysis, to name just a few [3, 25].

Although these examples illustrate considerably diverse applications and purposes of data mining, they all have one superior goal in common: The discovery of interesting, understandable and actionable knowledge from data. Nowadays, a broad variety of powerful and well-understood algorithms and tools are available for the discovery part of this task. However, most of the available technologies lack methods and user support for turning the mining results into domain knowledge. What Matheus et al. pointed out in 1993 [14], still holds today: Many data mining systems are great in deriving useful statistics and patterns from huge amounts of data, but they are not very smart in *interpreting* these results, which is crucial for turning them into *interesting, understandable and actionable knowledge*. Pattern analysis and post-processing often remains a laborious task for the user of a data mining system, assessing the interestingness and usefulness of discovered patterns is still considered being a hard problem [20].

We consider the lack of sophisticated tool support for incorporating human domain knowledge into the mining process as a main source of the shortcomings described above. We agree with Shen et al., who identified human intuition as the third crucial prerequisite for an effective discovery system besides inductive generation of hypotheses and their deductive verification [23]. Although the usefulness of domain knowledge exploitation for data mining has been widely recognized in recent years, it is still not fully understood and supported by mining tools. Although there exist numerous reports about successful exploitation of domain knowledge, especially for the data preparation and mining phases of the data mining process, few attention has been devoted to the questions of intelligent pattern analysis and how to integrate the discovered knowledge with the previously available one.

In our view, a next-generation data mining environment should actively support a user to both incorporate his domain knowledge into the mining process and update this domain knowledge with the mining results. Consequently, we consider a domain knowledge base a central component of future data mining environments. Inspired by the description of the 2001 workshop on "Integrating Data Mining and Knowledge Management" [11], we will deal with the questions regarding the integration of knowledge bases, mining tools and intelligent tools for pattern evaluation.

2 Interestingness, Beliefs and Domain Knowledge

Figure 1 shows a prototypical mining process derived from the CRISP-DM data mining methodology [4]. The upper arrows indicate feedback loops, the dashed arrows indicate (potential) flows of domain knowledge. The depicted process essentially is compatible to the classical KDD process model introduced by Fayyad et al. [7]. As can be seen from the figure, domain knowledge not only drives the initial phase of data min-

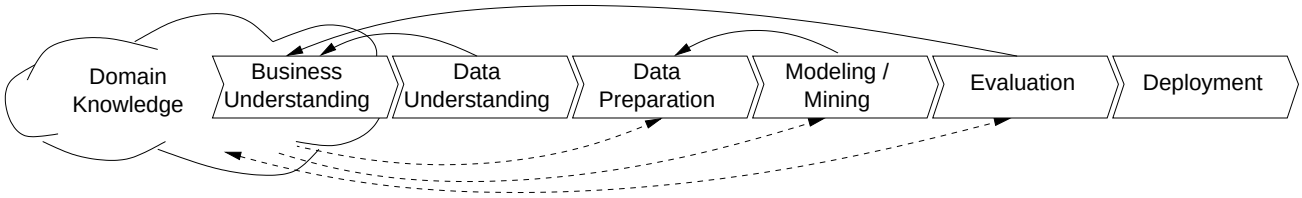

Figure 1: Domain knowledge and the mining process

ing, understanding the business (problem). It might also be applied for preparing and mining the data and clearly affects the analyst's satisfaction with the mining results as it ultimately determines the degree of their subjective interestingness. Furthermore, domain knowledge might in turn also be affected by the mining results, as these potentially lead to insights contradicting the initial knowledge and thus requiring feedback mechanisms for updating it.

As already denoted earlier, the ultimate goal of data mining is to find *interesting* patterns in data. The degree of a pattern's interestingness is assessed in the evaluation phase, which is the main focus of our work. In order to understand the role of domain knowledge in this context, we need to have a closer look on the concept of interestingness.

2.1 Interestingness

Literature distinguishes between *objective* and *subjective* measures of interestingness. Objective interestingness is generally based upon the inherent structure of mined patterns, i.e., the patterns' statistics like support or confidence. Patterns might be considered interesting when they represent strong regularities, rare exceptions, or when they help to distinguish different groups of items etc. In [19], Piatetsky-Shapiro has defined basic principles of objective rule interestingness. A comparison of a number of objective interestingness criteria can be found in [8]. Examples of objective interestingness of association rules have been presented by Tan et al. [27, 28]. Hilderman and Hamilton compare several statistical measures for ranking the results of online analytical processing applications in [9]. An innovative approach applying machine learning techniques for detecting interesting instances is proposed in [16].

An inherent weakness of these objective measures is caused by the fact that they do not make use of the human analyst's background knowledge about the application domain. We regard any knowledge used for data mining besides the data itself as domain knowledge, i.e. semantic meta-data, prior expectations and intuitions as well as any formalized or tacit knowledge about the application domain employed by an analyst in order to perform a data mining project. For example, a user might be able to distinguish interesting rare occurrences of a certain phenomenon from uninteresting statistical noise by resorting to background knowledge not available to the mining algorithm. This latter concept of interestingness is usually referred to as *subjective* interestingness, as it don't take only statistical properties of patterns into account, but also considers individual conditions of the respective human analyst.

An early attempt addressing this issue in KDD systems has been presented in [21]. A more general discussion is available with [24], where Silberschatz and Tuzhilin propose *unexpectedness* and *actionability* as user-oriented measures of pattern interestingness, which has been further elaborated in [18]. While the approaches relying on objective interestingness measures try to "guess" the subjective surprisingness from the information about a domain contained in the data itself ([8]), subjective measures rely on some formalization of expectations or previous knowledge. For example, in [18] Padmanabhan and Tuzhilin assume that initial beliefs about a domain have been generated by elicitation from a domain expert or learned from data. A recent overview of applications of soft computing technologies for the discovery of subjectively interesting rules and patterns is presented in [15].

2.2 The Role of Domain Knowledge

The literature already provides numerous examples of applications of subjective interestingness. In principle, most of them compare discovered patterns with some form of beliefs. Any mining results either supporting or contradicting these beliefs are considered being interesting. Those beliefs in turn are determined by domain knowledge. An analyst usually deduces his prior beliefs from the knowledge about a certain domain available to him. Hence, domain knowledge ultimately determines interestingness.

Yoon et al. categorize domain knowledge into the three types correlation, inter-field and category knowledge and propose appropriate acquisition and representation schemes for each of them ([29]). The authors then use this domain knowledge for mining query optimization. In [22], we have demonstrated how domain knowledge in the form of concept hierarchies can be used in order to improve Web mining results. In [2], Berendt and Spiliopoulou have shown how a similar approach can capture site semantics for intelligent Web mining. Liu et al. have presented their "Interestingness Analysis System" IAS, which comes close to our idea of an intelligent mining environment [12]. The authors require the user to express his/her various

types of existing knowledge in terms of a proprietary specification language. Based upon this knowledge base, the IAS aims at identifying conforming and unexpected rules returned by an association rule miner.

Although applied successfully with respect to their respective tasks, each of the above examples provides the drawback of requiring the establishment of proprietary knowledge bases. Furthermore, the user is encouraged to express domain knowledge in a very application specific form, scope and granularity, thus hindering the reuse of codified knowledge by other tools.

At the same time, research in the area of knowledge management has lead to quite mature standards for modeling and codifying knowledge. Today, ontologies are a key technology for intelligent knowledge processing, providing a framework for sharing conceptual models about domains [13]. There are already powerful tools available for capturing and management of ontological knowledge [5]. We propose to resort to exploit these advances in knowledge modeling for our purposes by applying them for the construction of intelligent data mining environments.

Early examples of successful implementations of this idea are already available: Hotho et al. use domain knowledge formalized as ontologies together with information extraction (IE) technologies for improving text mining algorithms and pattern interpretation. In [10], they discuss the application of codified background knowledge for different mining methods like document clustering, instance clustering and association rule mining. Their use of ontologies is manyfold: In data preprocessing, taxonomic ontologies are used for reducing the source data's dimensionality. For the mining phase, the authors exploit ontologies in order to improve clusterings and propose ontology-based similarity measures. Ontologies are also applied for mining generalized association rules. In post-processing, ontologies and their graphical representations facilitate pattern interpretation by the human expert [26].

2.3 Closing the Loop: Learning from Mined Knowledge

Once domain knowledge is properly integrated into the data mining process, it seems a natural consequence to search for ways of establishing feedback mechanisms that help to update the knowledge base once new and interesting insights about the domain have been discovered. As we propose to use standard knowledge representation mechanisms, this will help to reuse and disseminate the mined knowledge and keep the codified domain knowledge valid. Thus, knowledge base update mechanisms will be a major focus of our approach.

3 Design of a Domain Knowledge Enabled Mining Environment

When aiming at the integration of data mining and knowledge management solutions, we face three major challenges:

1. Mining tools represent discovered patterns in various forms and formats. In order to use these findings as input for domain knowledge-based pattern analysis, we need to transform these results into a sufficiently generic representation.

2. On the other side, we need mechanisms of deriving relevant beliefs for matching them with the discovered rules and patterns. These beliefs have to be expressed on the appropriate level of granularity.

3. As discussed in section 2.2, the subjective interestingness of discovered patterns is determined by supporting or conflicting beliefs. We need mechanisms to identify the conflict source and solving it appropriately.

Regarding the first challenge, there already exist several proposals which might evolve as future standards for describing statistical and data mining models, e.g. the Predictive Model Markup Language (PMML) [1]. One of our next tasks will be to evaluate these approaches with respect to our goal of developing a framework adoptable as widely as possible and practicable.

The second issue, deriving beliefs from the available knowledge, has basically two prerequisites: First, domain knowledge must be acquired from experts or other sources. A challenge here is that human knowledge is often tacit and imprecise/qualitative in nature, as compared to the data under analysis, which usually provides quite precise statistics. Second, the domain knowledge must be made available to the system in a representational form that allows for an efficient comparison with the discovered patterns. A major problem to be solved here is the fact that the patterns or rules present in the databases often refer to a different "level of knowledge" than the rules comprising the knowledge base. For example, a knowledge base about consumer behavior might state that beer is often bought together with diapers. This fairly general statement might be tested against patterns discovered from a data base containing sales data containing EAN codes instead of product categories. Here, ontologies not only provide a vehicle for the semantical information required to translate between different representational levels of granularity, but also for modeling horizontal relationships between concepts.[1]

[1] Research and practice provide numerous examples of resolving taxonomical issues (i) in data preprocessing or (ii) by mining algorithms. The first solution requires highly problem specific

The third challenge listed above refers to testing if patterns and rules discovered by data mining are consistent with the domain knowledge previously available. Generally, conflicts can occur due to a) unexpected patterns, b) violation of beliefs or c) violation of codified domain knowledge. Case a) refers to the discovery of entirely new knowledge, e.g. a shift in consumer preferences condensing in shopping basket data. Case b) might occur when prior assumptions codified in the knowledge base are incorrect. For example, when analyzing business processes, the average task completion time at a certain node in a process chain might be underestimated. The third case refers to the situation where the conflicting results are (tacitly) known to the analyst but the corresponding knowledge is simply not stated or incorrectly represented in the knowledge base. We will develop methods for identifying these different kinds of inconsistencies and for modifying the knowledge base as necessary.

4 Current State and Outlook

The dissertation project described above is still in quite an early stage. The rough outline has been fixed and we already have an initial overview of major parts of related work. Further literature research is currently done in the areas of reasoning about ontologies and transformation of knowledge representations. As technical platform for domain knowledge acquisition and modeling, we plan to use the infrastructure provided by the Protégé-2000 framework [17]. Our reasoning mechanisms and interfaces for accessing mining results are planned to become plug-ins to the Protégé system.

One idea currently under consideration for bridging the above mentioned gap between qualitative and somewhat "intuitive" human knowledge and the quantitative results returned by mining algorithms is to apply results from fuzzy set theory. Allowing for fuzzy quantifications of knowledge expressed in a sentence like "a significant number of online shoppers chooses credit card as payment option" might help assessing the subjective interestingness of patterns. An overview of related work currently under evaluation is presented in [15]. Furthermore, we're actually evaluating domains like Web usage or business process analysis as areas for case studies that will be subject of our proof of concept. First presentable results are expected to be available by Q4 of the year 2003.

References

[1] Data mining group (DMG) homepage. `http://www.dmg.org`. accessed 2003-07-21.

[2] Bettina Berendt and M. Spiliopoulou. Analysing navigation behaviour in web sites integrating multiple information systems. *VLDB Journal: Special Issue on Databases and the Web*, 9(1):56–75, 2000.

[3] Michael J.A. Berry and Gordon Linoff. *Data Mining Techniques for Marketing, Sales and Customer Support*. John Wiley & Sons, 1997.

[4] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. *CRISP-DM 1.0 – Step-by-step data mining guide*. The CRISP-DM Consortium / SPSS Inc., availabe on `http://www.crisp-dm.org`, 2000.

[5] A. J. Duineveld, R. Stoter, M. R. Weiden, B. Kenepa, and V. R. Benjamins. Wondertools? a comparative study of ontological engineering tools. *International Journal of Human-Computer Studies*, 52(6):1111–1133, 2000.

[6] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge discovery and data mining: Towards a unifying framework. In E. Simoudis, J. Han, and U. Fayyad, editors, *Proceedings of KDD'96, Second International Conference on Knowledge Discovery and Data Mining*, pages 82–88, Menlo Park, CA, 1996. AAAI Press,.

[7] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34, November 1996.

[8] A.A. Freitas. On rule interestingness measures. *Knowledge-Based Systems*, 12(5-6):309–315, October 1999.

[9] R.J. Hilderman and H.J. Hamilton. Evaluation of interestingness measures for ranking discovered knowledge. In Williams G.J. Cheung, D. and Q. Li, editors, *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'01)*, Lecture Notes in Computer Science, pages 247–259, Hong Kong, April 2001. Springer-Verlag.

[10] Andreas Hotho, Alexander Maedche, Steffen Staab, and Valentin Zacharias. On knowledgeable unsupervised text mining. In *Proceedings of the DaimlerChrysler Workshop on Text Mining*, Ulm, April 26–27 2002. Springer, to appear.

[11] Franz J. Kurfess and Melanie Hilario, editors. *Integrating Data Mining and Knowledge Management, Workshop held in conjuction with ICDM'01: The 2001 IEEE International Conference on Data Mining*, San Jose, CA, November 2001. IEEE.

data preparation, i.e. the whole mining process has often to be re-initiated in the light of sometimes only slightly changing analysis goals. By our work, we hope to leverage this problem by making mining results more "reusable". The latter approach can be considered complementary of our goal of pattern interpretation.

[12] Bing Liu, Wynne Hsu, Shu Chen, and Yiming Ma. Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems*, 15(5):47–55, 2000.

[13] Alexander Maedche, Boris Motik, Ljiljana Stojanovic, Rudi Studer, and Raphael Volz. Ontologies for enterprise knowledge management. *IEEE Intelligent Systems*, 18(2):26–33, March/April 2003.

[14] Christopher J. Matheus, Philip K. Chan, and Gregory Piatetsky-Shapiro. Systems for knowledge discovery in databases. *IEEE Transactions On Knowledge And Data Engineering*, 5:903–913, 1993.

[15] Sushmita Mitra, Sankar K. Pal, and Pabitra Mitra. Data mining in soft computing framework: A survey. *IEEE Transactions on Neural Networks*, 13(1):3–14, January 2002.

[16] Katharina Morik. Detecting interesting instances. In David Hand, Niall Adams, and Richard Bolton, editors, *Proc. of ESF Exploratory Workshop on Pattern Detection and Discovery*, number 2447 in LNAI, pages 13–23, London, UK, September 2002. Springer.

[17] N. F. Noy, R. W. Fergerson, and M. A. Musen. The knowledge model of protege-2000: Combining interoperability and flexibility. In *Proc. of 2nd International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000)*, volume 1937 of *LNCS*, pages 17–32, Juan-les-Pins, France, 2000. Springer.

[18] Balaji Padmanabhan and Alexander Tuzhilin. A belief-driven method for discovering unexpected patterns. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining KDD 1998*, pages 94–100, August 1998.

[19] Gregory Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In G. Piatetsky-Shapiro and W.J. Frawley, editors, *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, Cambridge, MA, 1991.

[20] Gregory Piatetsky-Shapiro. Knowledge discovery in databases: 10 years after. *SIGKDD Explorations*, 1(2):59–61, January 2000.

[21] Gregory Piatetsky-Shapiro and Christopher J. Matheus. The interestingness of deviations. In *Proceedings of KDD-94: AAAI-94 Knowledge Discovery in Databases Workshop*, pages 25–36. AAAI Press, July 1994.

[22] Carsten Pohle and Myra Spiliopoulou. Building and exploiting ad hoc concept hierarchies for web log analysis. In Yahiko Kambayashi, Werner Winiwarter, and Masatoshi Arikawa, editors, *Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery, DaWaK 2002*, volume 2454 of *Lecture Notes in Computer Science*, pages 83–93, Aix en Provence, France, September 4–6 2002. Springer-Verlag.

[23] W. Shen, K. Ong, B. Mitbander, and C. Zaniolo. Metaqueries for data mining. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 375–398. AAAI/MIT Press, 1996.

[24] Avi Silberschatz and Alexander Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970–974, December 1996.

[25] Myra Spiliopoulou and Carsten Pohle. Data mining for measuring and improving the success of web sites. In Ronny Kohavi and Forest Provost, editors, *Journal of Data Mining and Knowledge Discovery, Special Issue on Applications of Data Mining to Electronic Commerce*, volume 5, pages 85–114. Kluwer Academic Publishers, January–April 2001.

[26] Steffen Staab, Christian Braun, Ilvio Bruder, Antje Düsterhöft, Andreas Heuer, Meike Klettke, Gunter Neumann, Bernd Prager, Jan Pretzel, Hans-Peter Schnurr, Rudi Studer, Hans Uszkoreit, and Burkhard Wrenger. GETESS - searching the web exploiting german texts. In *CIA'99 – Proceedings of the 3rd Workshop on Cooperative Information Agents*, LNCS 1652, pages 113–124, Upsala, Sweden, July 1999. Springer.

[27] P. Tan and Vipin Kumar. Interestingness measures for association patterns: A perspective. Technical Report TR00-036, Department of Computer Science, University of Minnesota, 2000.

[28] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Indirect association: Mining higher order dependencies in data. In *Proceeding of PDKK 2000*, pages 632–637, 2000.

[29] Suk-Chung Yoon, Lawrence J. Henschen, E. K. Park, and Sam Makki. Using domain knowledge in knowledge discovery. In *Proceedings of the Eighth International Conference on Information and Knowledge Management*, pages 243–250. ACM Press, 1999.