DECISION-MAKER-AWARE DESIGN OF DESCRIPTIVE DATA MINING

Benedikt Kaempgen
Karlsruhe Institute of Technology (AIFB)
Karlsruhe, Germany
benedikt.kaempgen@kit.edu

Florian Lemmerich
University of Würzburg
Department of Computer Science VI
Würzburg, Germany
lemmerich@informatik.uni-wuerzburg.de

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

ABSTRACT

This paper presents two real-world case studies focussing on descriptive data mining for decision-makers. For that, we first propose a process-oriented design of descriptive data mining that helps in describing and performing such projects. Finally, we discuss important lessons learned during the implementation of the respective projects.

1. INTRODUCTION

With the implementation and collection of data in routine fashion, e.g., in industrial, medical, administrative and social-web-based scenarios, the analysis and mining of such accumulated data is of prime importance for intelligent decision support. However, currently up to 60% [1] of data mining projects fail. One problem concerns the integration of the key stakeholders in data mining projects, i.e., the decision-makers. They need to be tightly integrated into the project, similar to the actual data mining engineers. Thus, in order to improve the common understanding on goal, approach and outcome a more transparent data mining process considering both developer team and decision-maker is rather important.

In this paper, we consider two case studies: The first one is concerned with the analysis of the success and failures of (bachelor) student groups in order to help decision support for improving the success rate of individual curricula. The second one is concerned with the evaluation of a web-based training system and aims, e.g., at analyzing the outcomes of different study groups and their learning differences.

We focus on approaches for obtaining descriptive reports and descriptive data mining models, e.g., local patterns and rules as actionable knowledge for decision support. Descriptive data mining focuses on describing the data by the discovered patterns and relations: In contrast to predictive data mining no specialized model is extracted (for later prediction or classification) but a set of patterns and/or relations is mined for characterizing and describing the data and its hidden components.

In this context, the contribution of this work is threefold: First, we propose a process-oriented design for describing and performing projects in the context of decision-maker-aware descriptive data mining. Second, since only few descriptions of successful data mining projects that concentrate on decision-makers as well as the development team are available, we present two such case studies. Third, we discuss specific experiences and lessons learned during the implementation of the case studies. Altogether, it is our motivation to enable more successful descriptive data mining projects.

The rest of the paper is structured as follows: Section 2 discusses related approaches. After that, Section 3 presents the process-oriented design for describing and performing the case studies. Next, the implemented case studies are described in detail. Section 4 reports specific experiences and lessons learned obtained during the implementation of the case studies. Finally, Section 5 concludes the paper with a summary and interesting directions for future work.

2. RELATED WORK

In the following, we describe related work that deals with data mining design and implementations.

Process models provide an high level overview of the input and output of required data mining tasks. According to Kurgan and Musilek [2] CRISP-DM [3] is most prominently used in data mining projects. It consists of six iteratively executed phases: *Business Understanding* and *Data Understanding* make sure that the developer team has necessary background knowledge to deal with the problem of the decision-maker. In *Data Preparation* the available data is transformed for analysis, e.g., by selection, cleaning, construction, transformation and integration. In the *Modeling* step data mining techniques (algorithms) are applied to the prepared data to extract information and knowledge. In the *Evaluation* these results are evaluated, validated and checked against the data mining objectives. Finally, in the *Deployment* phase the results are employed for action, i.e., integrated into the respective processes of the decision-maker.

Marbán et al. [1] discuss the evolvement of data mining to an engineering discipline. They emphasize, that successful projects take more than CRISP-DM's Development Processes: *Organizational Processes* influence the whole organization in which data mining techniques are being used, e.g., continuous improvement and training or establishing of an appropriate data mining infrastructure. *Project Management Processes* assure successful project planing, e.g., by continuous communication with the decision-maker. Furthermore, *Integral Processes* support the development, e.g., documentation or configuration management. Although process models help developer teams and decision-makers to understand what to do in data mining projects, they do not describe *how* it can be done.

In contrast, *methodologies*, e.g., Catalyst [4] feature step-by-step guidance to data mining. However, as methodologies are more dependent on current techniques and systems, they are difficult to keep up to date.

Most *case studies* describe how techniques and systems can be applied in a specific project and concrete application domain. However, while many case studies of data mining projects have been presented (e.g., [5]), they are primarily used for demonstration of specific tools, results or techniques and therefore are seldom more generally applicable.

3. CASE STUDIES

In this section, we present two case studies. After presenting the process-oriented design, we discuss each one in detail.

3.1. Process-Oriented Design

Following Yin's [6] recommendations for well-designed case studies the purpose of the covered case studies is thoroughly describing how descriptive data mining can be successfully applied. As such the case studies are aimed at readers with both some technical background and business interest that consider data mining techniques in a project.

3.1.1. Focused Roles

On the one hand the decision-maker intends to benefit from data mining techniques. More precisely, the decision-maker has access to raw data and expects descriptive data mining techniques to extract information suitable to support his decision(s). The needs of the decision-maker are formalized as *requirements*.

On the other hand, the team of developers intends to fulfill the specified requirements by applying descriptive data mining tasks. The team usually consists of three kinds of experts [7]: *Data mining experts* are familiar with data mining techniques and the respective tools. *Data experts* offer thorough understanding of available and useful data, e.g., the data representation or the data acquisition process, while *domain experts* hold knowledge of the application area.

3.1.2. Focused Processes

We focus on three components (see Figure 1 for an overview): First, decision-maker processes are mainly related to the decision-maker, considering his or her specific needs. They include project definition, engineering of data mining requirements and result presentation. Second, developer team processes deal with techniques and systems that enable the developer team to fulfill the requirements and obtain useful results. Third, organization processes cover functions shared by different projects.

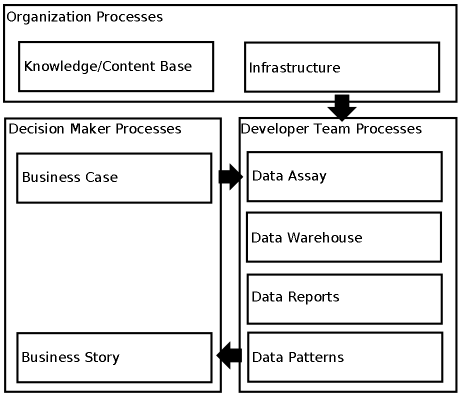

Fig. 1. Case Study Design w/ Information Flow

Decision Maker Processes Based on interviews with the decision-maker and possible feasibility studies, the developer team proposes a data mining approach to the decision-maker's problem in a *Business Case* document written "in management terms" [4, p. 205] and asks for his approval. The Business Case is a central document for any data mining project. It should include the background and motivation of the project, an explicit statement of the problem tackled by the project, a detailed description of the current situation and available data, recommended and alternative solutions, a project plan with time and cost estimations and a glossary.

As decision-maker and developer team mostly have different backgrounds, exact specification of suitable project requirements is a tedious, however, an essential task in descriptive data mining [8].

For that, the problem is restated in single "reporting type questions" [9] asking for attribute-value-pairs in tabular form describing instances of an object. These single Data Reports are then possibly analyzed further by "deeper analytic questions" [9] asking for hidden *Data Patterns* retrieved by techniques ranging from simple visualizations with diagrams or charts up to clustering or classification by machine learning algorithms.

To improve the decision-maker's understanding of the requirements both Data Reports and Patterns may be illustrated by (fictional) examples. Additionally, possibilities for evaluation might be given, e.g., background information and other (secondary) data.

A Business Case is not a static document. In fact, especially requirements will be exposed to constant changes. These are mainly due to results from development processes and have strong influence on the life cycle of a data mining project. In a successful project each requirement is fulfilled and documented in a *Business Story* [4, p. 509].

Developer Team Processes By preparing a *Data Assay* [4, p. 278] Business Understanding, Data Understanding and Data Preparation from CRISP-DM are implemented. It involves a concise description of the raw data, that is made available in a precisely specified tabular form. Additionally, quality issues, for example missing values, should be mentioned explicitly.

Data Preparation is done by making all neccessary data available in a *Data Warehouse*. The team identifies objects, attributes and relationships within the raw data and integrates them in an entity relationship model. Furthermore, data cubes are developed as a more subject-oriented view, if required. Each cell within a data cube can be described by shared attributes (dimensions) and aggregated attributes (measures). From these data cubes, a multidimensional model [10] is developed.

Next, the team creates *Data Reports*, which consist of a query from the data warehouse and additional layout information, e.g., a title or content explaining notes. Additional information can also be included as semantic annotations [11, 12], providing additional presentation possibilites and extended exchangeability. Based on these reports the team applies data mining algorithms to acquire *Data Patterns* specified in the requirements. Both data reports and mined patterns are evaluated and attached to the business story.

Organization Processes To support knowledge management between projects a standardized way of documentation is necessary. Instead of using single documents, we utilize a *Knowledge Base*, cf., [13], that supports references and more efficient searching. Based upon these approaches, we have designed an object-oriented documentation structure, that keeps track of various objects, e.g., goals, tasks, results, tools and documents, and their relationships, and makes these crucial experiences also available across different projects.

Also, a project can only be executed if an appropriate *Infrastructure* of hardware and software is available. For the different steps of our case study design highly specialized software components are available. For the Data Assay, for example, an ETL (Extraction, Transformation, Loading) component can be used, while implementing an entity relationship model or multidimensional model and or effective querying through SQL or MDX [1] is supported by specialized data warehouse components. A data reporting component makes it possible to customize data exports (CSV, ARFF) and to create reports with flexible layout information in various formats (e.g., PDF, XLS). A data mining component is able to read such exports and use data mining techniques (e.g., diagrams, correlation coefficients, subgroup discovery) on their data in order to make data patterns accessible. Finally, a documentation component supports web-based content management of objects, attributes and relationships.

The utilized documentation structure also provided the necessary information for an extensive description of the case studies.

3.2. Case Study I: Student Performance Evaluation

In the following, we describe the decision-maker processes, the developer team processes, and the organizational aspects of the bachelor project.

3.2.1. Decision Maker Processes

In Germany, the introduction of standardized bachelor degrees has been exposed to much criticism lately.

Therefore, for objective assessment on university level an in depth analysis is needed. Basic analytic questions to justify changes in the curriculum are for example: "How do important measures of bachelor degrees evolve?", "How do important measures of exams evolve?" or "What performance do current students achieve?".

The raw data for this proejct was provided by university administration. Since this data includes private student data, it was very carefully selected and precautiously pseudonymized. The legal process for getting permission to access the sensible data took several months in total. The data includes information on:

1. Enrollment information, with the actual semester, number of past semesters and degree of all bachelor students.

2. Exam information, with subject, number of achievable credits, number of lecture hours per week and the type of exam, e.g., module or submodule.

3. Information about student performance in an exam, with pass/fail status, achieved credits and mark.

4. Curricula information, that for each student separately defines categories to exams, e.g., obligatory or compulsory.

[1] http://msdn.microsoft.com/en-us/library/aa216767(SQL.80).aspx

Exemplary requirements, on which the head of the university faculty of (for example) biology, as a relevant decision-maker and the developer team might have agreed, is described as follows: As a Data Report, for each current student of biology the starting semester, number of past semesters, number of university semesters, sum of credits, average credits per semester and overall average grade should be presented. Additionally, the last two measures should be provided for each category of exam separately. As Data Patterns, for a better overview the reports were to be sorted on the number of past semesters and the sum of credits. Also, the histogram of credit points acquired by all students should be provided. This diagram was expected to reveal the number of very unsuccessful (and therefore probable to fail) and very successful (e.g. students already going to university before the end of college) students. Finally, student groups with low/high numbers of semesters and particlarly bad/low marks were to be discovered. This might extract information as "students in their fifth semester have an average mark of 2.0, students in their second semester have an average mark of 3.1, wheras all students have an average mark of 2.6".

During project life cycle these requirements were adapted several times. E.g., the formula for the computation of the overall average grade was not sufficiently specified at the project start. Furthermore, highly detailed requirements on the layout of result representations evolved. Since the utilized open source reporting software could not sufficiently support these requirements, tailored project specific java programs were additionally developed.

As part of the resulting business story the data report was given to the heads of faculities and provided insight into the overall student's performance. The credit distribution indicated a credit threshold for likely-to-fail-students suitable for an automatic warning system, that proposes these students for an additional mentoring program. Influences on student performance indicators will be further enhanced in the future with more information, e.g., survey answers, nationality, gender or age. Such reasons might propose actions towards a more adequate degree program. However, interpretations should be undertaken carefully. Students studying two-subject bachelor degrees need less credits in each subject and may indicate poor performance in comparison to others. Separating these student groups is issued to a follow up project.

3.2.2. Developer Team Processes

The developer team first imported several CSV file exports from the university information system into the data warehouse system. Based on that data, the team developed an entity relationship model made of five entities: Enrollment, person, exam, performance and exam category, each further described by attributes and relationships. Due to the complexity of SQL queries required for the data mining tasks, the ER-model was transformed into a multidimensional model. It contained two data cubes, one of enrollments and one of single performances.

Both an enrollment and a single performance are described by the student, the semester, the number of past semesters, the bachelor degree and an information whether that student is still enrolled in the actual semester. Each single performance is further described by the status, the exam and the type and category of the exam. For a data cell in the enrollment cube the number of individual students and both the minimal and maximal number of past semesters can be calculated. For a data cell of single performances the sum, number and average mark and the sum of credits can be calculated.

Now the team created reports based on data queries in MDX and specified layout informations according to the requirements. Additionally, exports for tools specialized on advanced pattern discovery were created. In this case distribution diagrams were created and subgroup discovery tasks were performed.

3.2.3. Organization Processes

As infrastructure three separate computer systems (each common 32-bit machines, 2 GHz, 2 GB RAM) were used: On one workstation the team mainly used Pentaho Data Integration[2] for the ETL processes and both VIKAMINE[3] and Weka[4] for data mining. On a server, MySQL and Pentaho Mondrian OLAP[5] were used for the data warehouse and Pentaho Business Intelligence Platform[6] was used for creating the data reports. As knowledge base the team used Semantic MediaWiki[7] on another server (for an overview, see Figure 2).

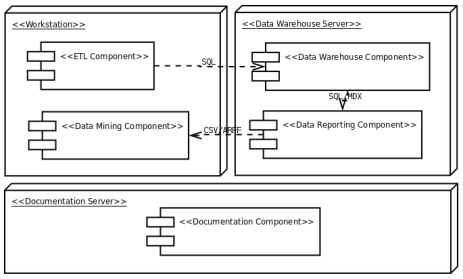

Fig. 2. Bachelor Infrastructure

The results of the project provided valueable insights on the performance of the students, on an automated and on-demand basis.

[2] http://kettle.pentaho.org/
[3] http://www.vikamine.org/
[4] http://www.cs.waikato.ac.nz/ml/weka/
[5] http://mondrian.pentaho.org/
[6] http://community.pentaho.com/projects/bi_platform/
[7] http://www.semantic-mediawiki.org/

3.3. Case Study II: E-Learning system evaluation

Again, the processes centric to the decision-maker, the developer team and the organization are discussed.

3.3.1. Decision Maker Processes

Students at the university of Wuerzburg are offered exam-relevant case-based training courses. The benefits of such a learning system need to be evaluated regularly. Exemplary questions include: "What influence does learning with the system have on exam performances?" or "How satisfied are users of the learning system?". User logs can provide useful data to answer such questions:

1. Log data tracks information about users learning with single cases. Each case execution consists of questions each offering a single score that is accumulated to a total score. The log data also contains information on the usage of help functions, e.g., asking for background information, reading hints or taking a break. Furthermore, at the end of most cases the user is asked for system evaluation: A mark about the case and the system and some textual feedback.

2. Meta information contains additional facts about cases: The form of case evaluation and the time the author expects a user to finish a case.

3. Exam results are available for some courses supported by case-based training.

Exemplary requirements can be described as follows: As a Data Report, for each exam result of a student the number of processed cases, the overall time used for learning with the system, the average overall practice score and the mark and percentage of correct answers in the exam are presented in tabular form. As Data Patterns, correlations between the engagement of the students with the system and their performances at the exam should be discovered, e.g., using a scatter plot and correlation coefficients. This requirement was initially expected to show a high influence of a student's effort with the system and his exam results, showing the effectiveness of the system. While providing promising results, however, no statistically significant correlation was discovered, in contrast to expectations: This is possibly due to not considered influences on student performances, e.g., present knowledge (level) of students, and due to a limited availability of (external) exam results in the considered sample of data.

3.3.2. Developer Team processes

The developer team first imported the provided data into the data warehouse system. This was a non-trivial task, since some data was available in a semi-structured form. Then, the team developed an entity-relationship model made of eight entities: student, case, case execution, evaluation, exam result, score, score action and case action. A multidimensional model consisting of three cubes was added for better querying. Each cube is described by several partially shared dimensions, e.g., student, case and date of execution. A case action is further described by the time of action (beginning and end of case execution) and the kind of action (e.g., pause, case summary, link). A case execution is further described by the exam that execution was relevant to. For a data cell of case execution actions the number and overall time of the actions can be calculated. For a data cell of case executions can be given e.g., the number of case executions, the average overall score, the overall time and the average performance of corresponding exams. For a data cell of scores the number of scores, the average score and the average/overall time taken for viewing the question and answer hints can be calculated. Similar to the bachelor case study, the developer team now designed data reports and exports as stated in the requirements, e.g., correlation mining.

3.3.3. Organization Processes

The Organization processes were executed similar to the bachelor case study. Both projects could not only use the same knowledge base but basically rely on the same infrastructure.

For examining the learning behavior of the students using the CaseTrain system, the performed reports and descriptive data mining results proved promising. Therefore, similar data mining approaches will be implemented as routine mechanisms within the CaseTrain system in the near future.

4. LESSONS LEARNED

From the case studies we could obtain several lessons learned: The proposed methodology appears to be generally applicable: Both projects – though substantially different in domain and requirements – were successfully finished; Data Reports in tabular form are flexible enough to contain most kinds of information; from simple diagrams to sophisticated machine learning algorithms – Data Patterns include the whole range of techniques to retrieve knowledge from this preprocessed raw data. Moreover, for most neccessary components open source software is available.

More than 70% of development time was used for the Data Assay and Data Warehouse. Changes to the data structure, e.g., when adding new features, result in significant additional work. Versionizing and refactoring of raw data description and preprocessing steps that get repeated several times would have been useful and seem essential in bigger projects.

Intensive documentation obviously is crucial for long-running data mining projects, especially if team members change. By documenting not only the project itself, but also sharing experiences and best practices, e.g., on applied tools and techniques, the documentation of one project proved to be extremely helpful for the other. Further cross-project benefits were achieved, since both projects shared a common infrastructure of hardware and software.

Legal aspects of a project should be addressed very early in a project, since the reviewing of data privacy issues and the integration of additional data can require a substantial amount of time. For having several and long running projects a framework of tools as used here seem crucial due to synergistic effects. The projects could be executed exclusively using open source systems. However, some components of current open-source system showed to be insufficient to match project requirements, e.g., highly specialized layouting of the results. Specifically tailored scripts were suitable to fill this gap. This combination of a tool suite for general purpose tasks and additional project specific implementations seems to be well suitable to handle highly specialized requirements.

5. CONCLUSIONS

This paper presented two case studies of successful descriptive data mining projects in two different contexts, i.e., the context of the analysis of university students performance and in usage data evaluation of an e-learning system. We proposed a decision-maker-aware approach for descriptive data mining, and discussed important lessons learned. In the future, in order to fully evaluate the decision-maker-awareness, retrieve general best practices and finally develop a full-scale methodology for descriptive data mining we aim to apply our design to further case studies in various domains.

6. ACKNOWLEDGEMENTS

Part of this work has been funded by the EU IST FP7 project ACTIVE under grant 215040, and by the German Research Council (DFG) under grant Pu 129/8-2. Furthermore, this work has been partially supported by the VENUS research cluster at the interdisciplinary Research Center for Information System Design (ITeG) at Kassel University.

7. REFERENCES

[1] Oscar Marbán, Javier Segovia, Ernestina Menasalvas, and Covadonga Fernández-Baizán, "Toward Data Mining Engineering: A Software Engineering Approach," *Information Systems*, vol. 34, no. 1, pp. 87 – 107, 2009.

[2] Lukasz A. Kurgan and Petr Musilek, "A Survey of Knowledge Discovery and Data Mining Process Models," *Knowl. Eng. Rev.*, vol. 21, no. 1, pp. 1–24, 2006.

[3] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth, "CRISP-DM 1.0 Step-by-step Data Mining Guide," Tech. Rep., The CRISP-DM consortium, August 2000.

[4] Dorian Pyle, *Business Modeling and Data Mining*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003.

[5] Michael Brydon and Andrew Gemino, "Classification Trees and Decision-Analytic Feedforward Control: A Case Study from the Video Game Industry," *Data Min. Knowl. Discov.*, vol. 17, no. 2, pp. 317–342, 2008.

[6] Robert K. Yin, *Case Study Research*, Number 5 in Applied social research methods series. Sage, Thousand Oaks, Calif. [u.a.], 4. ed. edition, 2009.

[7] Sarabot S. Anand and Alex G. Buchner, *Decision Support Using Data Mining*, Trans-Atlantic Publications, 1998.

[8] Paola Britos, Oscar Dieste, and Ramón García-Martínez, "Requirements Elicitation in Data Mining for Business Intelligence Projects," in *Advances in Information Systems Research, Education and Practice*. 2008, pp. 139–150, Springer Boston.

[9] Ron Kohavi, Llew Mason, Rajesh Parekh, and Zijian Zheng, "Lessons and Challenges from Mining Retail E-Commerce Data," *Mach. Learn.*, vol. 57, no. 1-2, pp. 83–113, 2004.

[10] Sergio Luján-Mora, Juan Trujillo, and Il-Yeol Song, "A UML profile for Multidimensional Modeling in Data Warehouses," *Data Knowl. Eng.*, vol. 59, no. 3, pp. 725–769, 2006.

[11] Martin Atzmueller, Fabian Haupt, Stephanie Beer, and Frank Puppe, "Knowta: Wiki-Enabled Social Tagging for Collaborative Knowledge and Experience Management," in *Proc. Intl. Workshop on Design, Evaluation and Refinement of Intelligent Systems (DERIS)*, 2009, vol. CEUR-WS.

[12] Martin Atzmueller, Florian Lemmerich, Jochen Reutelshoefer, and Frank Puppe, "Wiki-Enabled Semantic Data Mining - Task Design, Evaluation and Refinement," in *CEUR-WS 545*, 2009.

[13] Karin Becker and Cinara Ghedini, "A Documentation Infrastructure for the Management of Data Mining Projects," *Information & Software Technology*, vol. 47, no. 2, pp. 95–111, 2005.