Describing Decision Support, Data Mining, and Text/Web Mining Studies in SolEuNet

Marko Bohanec, Bojan Cestnik, Marko Grobelnik, Dunja Mladenić
Jožef Stefan Institute, Ljubljana, Slovenia,
Tel: +386 1 4773900; fax: +386 1 4251038
e-mail: marko.bohanec@ijs.si

Mário Alves, Alípio Jorge
LIACC, Oporto, Portugal

Steve Moyle
Oxford University, Oxford, UK

ABSTRACT

We present a schema for documenting and classifying completed Data Mining, Decision Support and Text and Web Mining cases. Project descriptions from these areas are unified in a hierachically structured relational database. The main objectives and benefits of the repository are presented and discussed.

1 INTRODUCTION

Working with end-user problems often implies that most of the results are confidential. They cannot be published even though the experts conducting the project have learned general lessons that can be potentially useful when approaching other end-user problems. That kind of experience is usually related to specific information about the problem characteristics and the used methodology. Usually, it can be shared without revealing confidential information about the problem and the customer.

In our work on developing prototype solutions for customer problems within project SolEuNet (Mladenić, 2001), we aim at solving end-user data mining, text/web mining and decision support problems (e.g., Cestnik and Bohanec, 2001), but also at developing new methods for collaborative data mining (Jorge, et al., 2002), combining problem solutions as well as combining data mining and decision support with information systems. The idea is to work on prototype solutions that have a potential for later commercial exploitation, and also to analyse failed and successful approaches using a joint infrastructure, education and dissemination. So, one of the main objectives is, based on the experience and lessons learned from practical cases, to propose a compact description of the cases in the form of a repository.

Among several benefits that are expected as a result of having the past projects stored in a repository, we emphasize the following ones:

- Unified project documentation;
- Stored knowledge and experience that could facilitate learning about the stored cases as well as replicating the successful solutions on similar new problems;
- Fast search among end-user projects by using descriptive criteria (assuming that the repository has been implemented in the form of a database);
- Summarized lessons learned from similar end-user problems, which might help avoiding obstacles when facing new problems.

The following section describes typical categories and examples of projects approached within SolEuNet. Section 4 then presents a unified project description schema, designed as a flexible relational data structure.

2 SolEuNet END-USER PROJECTS

End-user projects, approached within SolEuNet, belong to three different areas: (1) Decision Support, (2) Data Mining, and (3) Text and Web Mining.

Decision Support (DS). In SolEuNet, DS is mostly based on qualitative hierarchical multi-attribute modeling, using the supporting computer programs DEX and DEXi (Bohanec and Rajkovič, 1990; Bohanec, 2002). Seven different DS projects have been approached and completed. One of them, Housing (Bohanec et al., 2002), was aimed at supporting the task of housing loan allocation for the reconstruction of denationalised buildings in the city of Ljubljana. Two multi-attribute models have been developed and used for this purpose. The characteristics of this project – already using the unified description schema as proposed in section 3 – are shown in Table 1.

Prior to SolEuNet, completed DS projects had been documented in various ways. While some of them produced a written text report and/or some form of schematic description (Urbančič, et al., 1998), others were mostly documented with printouts from DEX and DEXi, and some outstanding projects were described as practical cases in scientific papers (e.g., Bohanec, et al., 1996).

Data Mining (DM). An example of a SolEuNet DM project is Mediana (Škrjanc, et al., 2001), where different data mining methods were used for the analysis of the media space in Slovenia. A media space consists of many different factors competing for the attention of the customer population in some environment. We have analyzed data describing the entire media space of the whole country (Slovenia) with the population of 2 million people. The data were collected by the private research institute Mediana. The database consists of 8000 questionnaires, each containing 1200 questions, gathered in 1998. The sample and the questionnaires were made by comparable research international standards.

Text and Web Mining. A SolEuNet problem of this kind comes from the Portuguese Institute of Statistics (INE), the governmental agency which is the keeper of national statistics. INE has the task of monitoring inflation, cost-of-living, demographic trends, and other important indicators. Its goal was to get information and on this basis provide better services on Infoline (www.ine.pt), a web site that makes statistical data available to the Portuguese citizens. The specific task was to extract knowledge from the web site's access data log, using DM techniques such as association rules, clustering and classification (Jorge and Moyle, 2002; Alves and Jorge, 2002). Association rules, for instance, can tell what is the next page a user would like to see, and help them finding the information they are looking for. This ability of "guessing" the user's wishes can be provided to the site by analyzing the usage of the site by other users, and discovering their own preferences. Also, the technique of clustering can, from the same stream of data, discover natural groups of users with similar preferences and behavior. This knowledge can help improve the usability of the site. Data collection is nearly costless, but the patterns found in the data can help the Portuguese save thousands of hours in their quest for statistical data.

Initially, several project description schemas for these specific areas have been designed by different SolEuNet workpackages (Mladenić, 2002). For instance, a description schema for DS projects was proposed in (Cestnik and Bohanec, 2002). A different schema was used for INE (Jorge and Moyle, 2002). Almost independently, the SolEuNet Information Collector (SENIC) database has been developed as a web system designed to support the task of collecting information about tools and case studies in SolEuNet. SENIC was engineered with the reliable web technologies described in (Alves, 2001). Although designed as a general repository, SENIC has been found more appropriate for describing DM than DS projects, clearly exposing the need for a unified project description schema.

3 UNIFIED PROJECT DESCRIPTION

The unified approach to describing Data, Text, and Web Mining and Decision Support solutions of completed end-user projects draws on two facts. First, these projects share a considerable number of common characteristics, which can be used for all of them. For example, all projects have descriptions such as title, keywords, summary, and data about the end-user. Second, project descriptors can be layered in order to cope with the specifics of approaches and applied methods in different areas.

This leads to a hierarchically organized relational database in which, at the top level, a project description is divided into three categories: (A) general description, (B) problem description and (C) method-specific parameters. This division is rather natural: first, a project is described in general, regardless of the specific type of the project and applied methods. Then, the specific problem is elaborated in more detail, using descriptors that are specific for the taken approach, such as DM or DS. Finally, method-specific parameters are presented on the third level.

Each higher-level category can contain one or more lower-level categories. For example, consider a hypothetical project, whose general characteristic can be described by descriptors of the category A. Suppose it is a DS project; in this case, the description can be supplemented by DS-specific parameters B. The problem can be approached by one or more different DS methods (C), for instance by two qualitative multi-attribute models (C1 and C2), a quantitative multi-attribute model (C3) and decision trees (C4). In addition, the same project (A) may have some data available, which can be analyzed by DM techniques and thus can be described by DM-specific parameters (say, B2). Again, several methods can be used for DM, such as association rules (B2.C1) and clustering (B2.C2).

Thus, this hypothetical project can be described by the following instance of the unified schema.

A:General description (Project acronym, Title, Keywords...)

- B1: DS Problem description: Background, Problem style, Evaluation
    - C1: First DS qualitative multi-attribute model
    - C2: First DS qualitative multi-attribute model
    - C3: DS quantitative multi-attribute model
    - C4: DS decision tree
- B2: DM Problem description: Background, Problem style, Evaluation
    - C1: DM association rules
    - C2: DM clustering

Organized in this way, the schema is highly flexible. First, it facilitates the description of projects that are approached by a variety of different approaches and methods. Second, it can be easily extended by new sets of descriptors corresponding to new types of problems (B) or new methods (C).

Table 1. Project Housing described by the unified schema.

A. General	
Project acronym	Housing

Project title		Loan allocation for the Housing Fund of Ljubljana
Keywords		Loan allocation, housing
Business sector		Finance
End-user mission		Housing, mortgage market
Customer institution		The Housing Fund of Ljubljana Municipality
Location		Ljubljana, Slovenia
Involved SolEuNet partners		Temida, IJS
Other partners		None
Start date		January 2000
End date		September 2001
Time span		9 months
Expert team size		5
Expert resources		14 MM
Press release		*text describing the project (omitted)*
Summary		Decision support of a tender for renovating old denationalized blocks of flats in Ljubljana
B. DS Problem Description		
Background	Problem acronym	Housing
	Problem title	Loan allocation
	Business success criteria	Undefined
	Internal champion	Not available
	Problem owner(s) accessible	Yes
Problem style	Problem type	Two-time
	Problem structure	Semi-structured
	Problem definition	Medium
	Organizational level	Tactical/strategic, management involved
	Supporting methods	Modelling, qualitative ranking/evaluation models, computational models, database, what-if analysis
	Primary DS elements	Data, models
	Group decision problem	No (no different interests)
Team members	Problem owner	1
	Additional experts	1
	Decision analysts	3
	Users	0
	Others	0

C. Method-specific parameters			
		C1.	**C2.**
Method type		Qualitative multi-attribute model	Qualitative multi-attribute model
Model name		A	B
Model description		Priority ranking of applicants that own only one flat in which they reside (the flat must be in a denationalised block)	Priority ranking of applicants that own aother denationalised flats rented non-profitably
Tools used		DEX	DEX
Size	Basic attributes	10	6
	Aggregate attributes	7	4
	Ranks	5	5
Number of options		109	258

Table 2. Project INE described by the unified schema.

A. General	
Project acronym	INE
Project title	Web access log analysis for INE
Keywords	Web access analysis, clustering, data mining
Business sector	Public agency
End-user mission	Compiler and keeper of the Official Statistics for Portugal
Customer institution	INE: Instituto Nacional de Estatística
Location	Porto, Portugal
Involved SolEuNet partners	LIACC, IJS, OFAI
Other partners	None
Start date	October 2000
End date	October 2002
Time span	25 months
Expert team size	6
Expert resources	22 MM
Press release	*text describing the project (omitted)*
Summary	Web access log analysis for the Portuguese Institute of Statistics, Porto (INE)

B. DM Problem Description		
Background	Problem acronym	INE
	Problem title	Log analysis
	Business success criteria	Undefined
	Internal champion	Available
	Problem owner(s) accessible	Yes
Problem style	Representation	Converted to relational data base
	Problem type	(1) Characterization, (2) Clustering, (3) Symbolic classification
	Problem definition	Broadly defined
Data	Number of tables	3+
	Number of attributes	32
	Number of records	86000
	Cell footprint	8256000
	Quality	Low
Evaluation	Human evaluation expertise available	Yes
	Outcome measure	None
	Validation possible	No
	Validation technique(s)	None

C. Method-specific parameters		
	C1.	**C2.**
Method type	Association rules (Apriori)	K-means clustering
Tools used	CLEMENTINE	CLEMENTINE
Number of models	3	3
Size of models	number of rules in [10, 20]	6 clusters
Parameter setting	minimum rule coverage = 5%; minimum rule accuracy = 60%; evaluation measure = difference of confidence quotient to 1; evaluation measure lower bound = 50%	K=6

For the illustration of specific descriptors, the DS project Housing is described by this schema in Table 1. Notice that the descriptors in section A are standardized and equal for all projects. Section B is specific to DS projects, but equal

for all of them. Section C contains two descriptions, C1 and C2, each corresponding to one of the multi-attribute models developed in the project.

For another example, Table 2 presents the description of the INE project. Notice that the same project descriptors as in Table 1 are used in part A. However, INE is a DM project, not a DS one as Housing, so the two tables differ in parts B and C. Table 2 contains descriptors applicable to DM problems (part B) and two specific DM methods (parts C1 and C2).

4 CONCLUSIONS AND FURTHER WORK

The main goal of this work was to propose a unified schematic description of completed end-user cases that can serve as a basis for the repository. The repository is one of the prerequisites for promoting and extending exploitation of Data Mining, Decision Support and Web/Text Mining technology into practice.

There are several benefits of having the past projects stored in a repository. First, the stored projects are documented in a similar formal way; as a result, it is relatively easy to get information about a single project as well as to mutually compare two or more projects. Second, stored knowledge and experience in the repository facilitate the discovery and learning about the recorded cases as well as replicating the successful solutions in similar new problems. Next, when the repository gets implemented in the form of a database, it will facilitate fast searching among the stored projects by using descriptive criteria. Last but not least, one can gain access to summarised lessons learned from similar problems, which might help avoiding obstacles when facing new problems.

The proposed project description schema is highly flexible. Its hierarchical structure facilitates the description of problems that are of different types and that are approached by a variety of methods. Also, it can be easily extended to new types of problems and methods used.

For further work we plan to implement the resulting repository schema as an object-oriented computer database, accessible through WWW, and include additional completed projects in the repository.

5 ACKNOWLEDGEMENT

The work reported here was in part supported by EU project SolEuNet, IST-1999-11495, and by the Slovenian Ministry of Education, Science and Sport.

References

Alves, M.A.: Safe Web Forms and XML Processing in Ada. In: Reliable Software Technologies: Ada-Europe 2001: Leuven, Belgium, May 14-18. Springer, LNCS 2043. 349–358.

Alves, M.A., and Jorge, A.: INE's Infoline Website Access Analysis (www.ine.pt) : CRIPS-DM Report. February 2002. SolEuNet working document published on Zeno (http://zeno.gmd.de/login) /SolEuNet /WP5 /RAMSYS /INE_Infoline /Phases (restricted).

Bohanec, M., Rajkovič, V.: DEX: An expert system shell for decision support. Sistemica, 1(1), (1990) 145–157.

Bohanec, M., Cestnik, B., Rajkovič, V.: A management decision support system for allocating housing loans. In: Humphreys, P, Bannon, L., McCosh, A., Migliarese, P., Pomerol, J.-C.(eds.): *Implementing Systems for Supporting Management Decisions*.London:Chapman and Hall (1996).

Bohanec, M.: DEX: An expert system shell for decision support. `http://www-ai.ijs.si /MarkoBohanec/dex.html` (2002).

Bohanec, M., Rajkovič, V., Cestnik, B.: *Report on Decision Support Practical Cases, Phase II*, Jožef Stefan Institute, Ljubljana, Report DP-8512 (2002).

Cestnik, B., Bohanec, M.: Decision support in housing loan allocation: A case study. In: Giraud-Carrier, C., Lavrač, N., Moyle, S., Kavšek, B. (eds.): *IDDM-2001: ECML/PKDD-2001 Workshop Integrating Aspects of Data Mining, Decision Support and Meta-Learning: Positions, Developments and Future Directions*, Freiburg (2001) 21–30.

Cestnik, B., Bohanec, M.: *SolEuNet Report on the repository of problem descriptions:D6.6* (2002).

Jorge, A., Moyle, S.: *Sol-Eu-Net WP5 Data Mining midterm* report: D5.3.2 (2002).

Jorge, A., Moyle, S., Voss, A.: Remote Collaborative Data Mining Through Online Knowledge Sharing. In: Camarinha-Matos, L.M. (ed.): Collaborative Business Ecosystems and Virtual Enterprises. Kluwer Academic Publishers, 2002.

Mladenić, D.: EU project: Data mining and decision support for business competitiveness: a European virtual enterprise (Sol-Eu-Net). In: D'Atri, A., Solvberg, A., Willcocks, L. (eds.). *OES-SEO 2001: Open enterprise solutions: Systems, experiences and organizations*. Rome, 14-15 September 2001. Roma: LUISS (2001) 172–173.

Mladenić, D.: Describing Data Mining and Decision Support Studies in SolEuNet. *Technical Report IJS-DP 8622,* J.Stefan Institute, Ljubljana, Slovenia, May 2002.

Škrjanc, M., Grobelnik, M., Zupanič, D.: Insights offered by data-mining when analyzing media space data. *Informatica* 25(3), (2001) 357–363.

Urbančič, T., Križman, V., Kononenko, I: *Review of AI Applications*, Jožef Stefan Institute, Ljubljana, Report DP-7806 (1998).