THE Sol-Eu-Net PROJECT
DATA MINING LESSONS LEARNED

Dunja Mladenič, Nada Lavrač, Peter Ljubič,
Branko Kavšek, Marko Grobelnik
Department of Intelligent Systems, Jozef Stefan Institute,
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773377; fax: +386 1 4251038
e-mail: dunja.mladenic@ijs.si

ABSTRACT

This paper reports on data mining experiences of the 5th Framework project Data Mining and Decision Support for Business Competitiveness: A European Virtual Enterprise (Sol-Eu-Net). The data mining lessons learned are reported from the following perspectives: application results, business, views of Sol-Eu-Net partners acquired by interview technique, and lessons learned in two particular data mining projects: analysis of Web education materials and UK traffic accident data analysis.

1 INTRODUCTION

Most journal, conference and workshop paper report on new developments – new techniques, tools and applications – showing only the positive aspects of such developments. In brief, papers are full success stories, yet reported failures, steps leading to success, steps that failed, and representation choices that were critical to success are crucial for increasing the awareness of which approach to use in which situations. In addition, papers often describe successful applications, but they rarely include expert evaluations of results.

The purpose of this paper is to report on data mining experiences of the 5th Framework project Data Mining and Decision Support for Business Competitiveness: A European Virtual Enterprise (Sol-Eu-Net, IST-1999-11495, 2000-2003) [1], aimed at gathering the project data mining lessons learned. Our ultimate goal is to learn how to develop better applications from successes (evaluated by domain experts) and, most importantly, from failed data mining endeavors (why was an approach not successful, what can dataminers learn from this experience).

The Sol-EU-Net data mining lessons learned are reported from the following perspectives: application results (Section 2), business (Section 3), views of Sol-Eu-Net partners acquired by interview technique (Section 4) and particular lessons gathered in two particular data mining projects: analysis of Web education materials (Section 5) and UK traffic accident data analysis (Section 6). We conclude with the summary of the Sol-Eu-Net data mining project experience.

2 Sol-Eu-Net DATA MINING APPLICATIONS

Applications addressed by the Sol-Eu-Net project consortium are listed below:

- Analysis of media research data for the MEDIANA marketing research company [2].
- Brand name recognition for a direct marketing campaign for the Kline&Kline marketing company [3].
- Customer quality evaluation and stock market prediction for a large Australian financial house [4].
- Predicting the use of resources in Czech health farms [5].
- Analysis of patient groups at risk for coronary heart disease [6].
- Analysis of Web page access to improve site usability for a Portuguese statistics institute [7, 8].
- Analysis of IT projects funded by the European Commission [9].
- Analysis of international building construction projects [10].
- Automatic ontology construction from education materials on the Web for a large Slovenian publishing house [11].
- Analysis of UK road traffic accident data [12].

Data mining lessons learned from the two last-listed applications are analysed in Sections 5 and 6.

3 BUSINESS LESSONS LEARNED

Approaching customers in need of Data Mining solutions is usually non-trivial. The first issue is finding a common language that is time consuming but necessary. We need to get some idea about the business area and in order to identify a problem domain, the customer needs to have some idea about relevant capabilities of Data Mining methods. These initial, time consuming efforts may result with understanding that currently there is no intersection between the customers needs and the available methods. Unfortunately, this is a very common situation, mainly due to unsufficient understanding of the methodology potentials and requirements on the side of potential customers. That is why educating customers is so important and helpful in identifying potentials for our methods. In that direction presenting right level of details is an important issue. We do want to give simple explanations of the Data Mining methodology but also to show its potentials.

We have found that it is very helpful to have an inside enthusiast – a person at the customers site seeing potential of our methods, for instance in our application of Web log analysis for Portuguese statistical office [7, 8]. In general our experience is that working with customers needing Data Mining methods is demanding but a lot of fun, provides new ideas and motivation for further research.

4 Sol-Eu-Net LESSONS LEARNED FROM DATA MINING APPLICATIONS

This section reports on views of Sol-Eu-Net lessons learned as perceieved by the project partners. Their views were acquired by means of interviews.

Sol-Eu-Net data mining projects are intended to be of a collaborative nature. This means that several dislocated partners (typically across Europe) work together on the same project. This nature of projects causes some serious problems. First is due to the lack of communication. In the course of the Sol-Eu-Net project duration we have realized that it is crucial that people working on the same project get to know each other. It is also important to find out what methods they use to obtain the results. This problem can be partly aleviated by means of a kick-off workshops (e.g., collaborative data mining of UK traffic accident data kick-off meeting in Bristol in June 2002, see Section 6). At such workshops, data miners meet with end-users and come to know their needs, goals, their philosophy. At the Bristol two-day workshop, everybody's opinion was that it should have lasted at least one week. From the data understanding point of view, the meeting was long enough, but it was too short for sharing and proposing new ideas, and for actually doing the initial collaborative work.

Later, when a project is already 'alive', the communication is still a problem, even though the collaboration of project partners in data mining applications is performed by using a workgroup support system (system ZENO [13] is used in Sol-Eu-Net). The system is used to share tasks, ideas, results, and comments to some extent. Why is it not regularly used? A common opinion is that researchers do not really want to publish 'failures', even though this would be useful for the overall success of a project. So they try harder to find excellent solutions, which takes time, and in this way the whole idea of collaboration fails. What also happens is that researchers don't publish results at all. When finally results are published, it is usually not in a preferable format, since every researcher uses different tools. Standards would improve efficiency and cut-off overhead. More workshops in later phases of projects would increase productivity (similar to weekly meetings in corporations). Of course, due to large costs, such workshops would not be organized every week.

Another problem is the lack of data management. Project partners prepare data themselves, usually there are at least as many preprocessed data sets and data formats as the number of collaborating partners. There is no uniformity of data, so it is very hard or at least very time-consuming to reproduce the results. Some data miners suggested the use of a centralized database, which would be managed by a single administrator. Others see the solution in collaborative data preprocessing. Some problems were addressed by the use of the SumatraTT [14] data preprocessing tool, which, however, did not satisfy most of project partners.

From the data mining project management point of view, it is important to have a strong and committed lead person, what was not the case in most Sol-Eu-Net data mining projects. Project leaders should insist that researchers share partial results and ideas. A lead person should master information management and if a lead person is one of the researchers, the time he/she is willing to spends on managing the customer project is restricted and usually not in the main stream of the persons efforts.

Besides the above-listed main problems emerging in collaborative projects, there are also more pleasant lessons learned. People learned about others' approaches to solving problems. People respected and understood different cultures, habits etc., which is a success of its own, which is important for the future collaboration of experts from different institutions.

5 PUBLISHING HOUSE DATA MINING LESSONS LEARNED

One of the Data/Text Mining problems addressed in the Sol-Eu-Net project was for the biggest Slovenian publishing house. They were interested in text search with several non-standard functions and in automatic document categorization. The collection they have consists from text documents giving educational materials for different areas and different levels of primary, secondary and high school education. Materials were prepared with contractual authors, mostly distinguished authorities from the specific field and then edited by either in-house editors or in cooperation with other experts for pedagogy, methodology and didactic. As different authors have different possibilities and preferences when working on computers, material were potentially provided in different formats and transformed only as needed for the "classical paper-publishing" procedure. There was an ongoing project at the customers site on developing uniformly formatted database of educational materials based on uniform ontologies with the future goal of offering access on electronic media (CDs and Web) as well as on printed materials as well as on printed materials and to automate the process for the selection and categorization of educational material

This cooperation with publishing house gave us two kinds of experience: (1) motivation for new technical solutions and (2) specifics on cooperation with the people working in the publishing house. First experience is in meeting very unclean text data edited by several editors, existing in several formats (XML, HTML, ordinary text). Since the requirements were to build several solutions in the areas of data search engine and construction of hierarchical index, we built as many customized filters, as there were data formats. This solution enabled us to have text in the same format. The second experience is about requirements for customized search engine. Usual requirements include simple query language with basic logical operators (AND, OR, NOT). In our case the publishing house required special type of query operators since the system was used for the special type of data (history). As a result we added a mechanism for inexpensive inclusion of new complex query operators at the expense of additional space. Next, an interesting technical solution was found for the problem of hierarchical index construction, where the hierarchy was given in advance with no additional descriptions. We made simple approximate join between given hierarchy and documents in the database using search engine, which produced very useful results. For the second problem when making the hierarchy out of the set of documents, we used relatively novel approach using 2-means clustering, dividing the set of documents into two groups, and further dividing each of the group into new subgroups etc. since some intuitive the stopping condition was met. This procedure constructed very interesting hierarchy of structured subjects, which can serve as a good starting point for human editors.

We have found that Text Mining seems very interesting and promising for the content based industries like publishing and media houses or companies dealing with large amounts of documents. There are two major aspects the publisher found interesting and promising. First is that methods and tools based on Data and Text Mining add value to on-line services. Personalisation of learning, ontology based dynamic catalogs of content, extensive search and find methods as well as group work are already becoming distance learning system necessities. Second finding is that with the help of Data and Text Mining tools several activities in the publishing process could be optimised and automatised. Automatic categorization makes in the world of information overload the basic publishing process of "knowledge refinery" realistic and manageable. Extensive searchers help editors to overcome the "click-and-miss" stresses in daily work. So in general the filling from the publisher about these kind of technology was enthusiastic, positive and very promising.

But before they came to that conclusions the experience from our site was that the solutions provided by Text Mining technology seemed rather new for this type of customers where the most common practice is two level manual work (editing and design) with not enough experience in on-line contents. This led us to long initial phase of persuasion of the customer about the benefits of new approaches work the business. For that reason we think that the most convenient way to work with customers is to first educate somebody within the customer company providing some general, conceptual knowledge about Data and Text Mining methods.

6 UK TRAFFIC ACCIDENT DATA MINING LESSONS LEARNED

Another problem addressed by the Sol-Eu-Net project was the UK traffic accident data mining problem, where the end-user was interested in getting some useful information from a “relational” data set of all accidents that happened in the UK over a period of 20 years (from 1979 to 1999). The goal/challenge for the participants in the Sol-Eu-Net project was to try a variety of data mining techniques and share the produced knowledge to achieve better results. Additionally, constant collaboration with the end-user was crucial here for interpreting the results.

Due to the collaborative way of the project, several problems arise. First of them is the lack of communication. On the initial kick-off meeting with the end-user there was a lot of enthusiasm among the project partners. A lot has been said on how to collaborate, how to share data and results and how to use the workgroup support system ZENO [13] with its RAMSYS extension to do this. After the kick-off meeting, when all the good words and enthusiasm should be put to action, the communication and collaboration dropped partially due to the summer conference season and vacations. This could also be attributed partly to a relatively short kick-off meeting, which should have lasted at least one week, giving a chance to the partners of the project to define more in detail the process of information sharing. Partly the lack on more intensive communication is also due to the inherent resistance of the researchers to share experience on failed approaches, quietly working till they get very good results.

Another problem that showed up was lack of standards to share the data as well as the results. For data sharing a central database was suggested, but sharing results proved more difficult. The most difficult however is sharing knowledge.

In spite of the problems that had to be dealt with, there were also positive aspects of the collaboration, the most important being the kick-off meeting with the end-user. At such meetings, all the partners get to know each other and usually a great part of the work is done there. Another positive aspect is that partners in the project became aware of knowledge pooling and became accustomed to workgroup support systems. Otherwise, a known problem of lack of trust between business and academia was not an issue in this project, because of a very “disposed” end-user which knew how to listen to the academic arguments. He was also a very active and responsive member of the whole Traffic project team.

7 FINAL REMARKS

The data mining lessons learned are reported from the following perspectives: application results, business, views of Sol-Eu-Net partners acquired by interview technique, and lessons learned in two particular data mining projects analysis of Web education materials and UK traffic accident data analysis.

References

[1] Mladenić D. (2001). EU project: Data mining and decision support for business competitiveness: a European virtual enterprise (Sol-Eu-Net). In *OES-SEO 2001: Open enterprise solutions: Systems, experiences and organizations*. Roma: LUISS, pp. 172–173.

[2] Škrjanc M., Grobelnik M., Zupanič D. (2001). Insights offered by data-mining when analyzing media space data. *Informatica (Ljublj.)*, vol. 25, no. 3, pp. 357-363.

[3] Flach P.A., Gamberger D. (2001). Subgroup evaluation and decision support for a direct mailing marketing problem. In *ECML/PKDD'01 Workshop notes on Integrating Aspects of Data Mining, Decision Support and Meta-Learning*, pp. 45-56.

[4] Škrjanc M. (2000) Summary report on the analysis of financial data (confidential), Internal IJS Report.

[5] Stepankova O., Klema J., Lauryn S., Miksovsky P., Novakova L. (2002). Data Mining for Resource Allocation: A Case Study. In *BASYS-2002*.

[6] Gamberger D., Lavrač N. (2002). Descriptive induction through subgroup discovery : a case study in a medical domain. In *Machine learnin: proceedings of the Nineteenth International Conference* (ICML 2002), University of New South Wales, Sydney, Australia. San Francisco: Morgan Kaufmann, pp. 163-170.

[7] Grobelnik M., Mladenić D., (2002). Efficient Visualization of Large Text Corpora, *The 7th TELRI Seminar "Information in Corpora"*.

[8] Alípio J., Mário A.A. (2001). End of Phase I summary on INE Infoline: a Sol-Eu-Net end-user Phase I project, *IST-1999-11495 Project Progress Report* (internal).

[9] Grobelnik M., Mladenić D., (2002). Collaborative approach to Web access analysis. *International Conference on Methodology and Statistics*, Ljubljana.

[10] Moyle S., Bohanec M., Ostrowski E. (2002). Large and Tall Buildings: A case study in the application of Decision Support and Data Mining In *Proceedings of ECML/PKDD-2002 Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning*.

[11] Grobelnik M., Mladenić D., Jermol M. (2002). Exploiting text mining in publishing and education. In *Proceedings of the 5th International Multi-conference IS-2002*, Ljubljana: Institut Jožef Stefan.

[12] Flach, P., (2002) UK Traffic problem, Internal report.

[13] Voss A., Richter G., Moyle S., Jorge A. (2001). Collaboration support for virtual data mining enterprises. In *Proceedings of the 3rd International Workshop on Learning Software Organizations* (LSO'01), volume 2176 of Lecture Notes in Computer Science, pp. 83--95. Springer-Verlag.

[14] Aubrecht P., Železný F., Mikšovský P., Štěpánková O. (2001). Progress in SumatraTT: ILP connectivity and more new features. In *Proceedings of the 4th International Multi-conference IS-2001*, Ljubljana: Institut Jožef Stefan.