An architecture for an effective usage of data mining in business intelligence systems

Ana Azevedo, ISCAP/IPP, Porto, Portugal, aazevedo@iscap.ipp.pt
Manuel Filipe Santos, Department of Information Systems, University of Minho, Portugal, mfs@dsi.uminho.pt

Abstract

Business Intelligence (BI) is one emergent area of the Decision Support Systems (DSS) discipline. Over the last years, the evolution in this area has been considerable. Similarly, in the last years, there has been a huge growth and consolidation of the Data Mining (DM) field. DM is being used with success in BI systems, but a truly DM integration with BI is lacking. Therefore, a lack of an effective usage of DM in BI can be found in some BI systems. An architecture that pretends to conduct to an effective usage of DM in BI is presented.

Key words: Business Intelligence, Knowledge Discovery on Databases, Integration, Data Mining Languages

Introduction

Business intelligence (BI) can be defined as the process that transforms data into information and then into knowledge (Golfarelli *et al.*, 2004). Being rooted in the decision support systems (DSS) discipline, BI has suffered a considerable evolution over the years and is, nowadays, an area of DSS that attracts a great deal of interest from both the industry and researchers (Arnott & Pervan, 2008; Clark *et al.*, 2007; Hannula & Pirttimäki, 2003; Hoffman, 2009; Negash, 2004; Richardson *et al.*, 2008; Richardson *et al.*, 2009).
Data mining (DM) is being applied with success in BI and several examples of applications can be found. Despite that, DM has not yet reached to non specialized users. The authors consider that the real issue is related with the fact that knowledge Discovery in Databases (KDD), as presented by (Fayyad *et al.*, 1996), is not fully integrated in BI and DM systems are functioning as separate isles. Consequently, its full potential could be not completely explored by decision makers using the system. The authors consider also that one important issue is that it has not been given sufficient emphasis to the development of solutions that allow the specification of DM problems through business oriented languages, and that are also oriented for BI activities.
This paper presents an architecture that conducts to an effective usage of DM in BI systems. The utilization of a DM language that pretends to be business oriented and that allows an effective usage from business specialists is introduced. The main contribution of this paper is to present the viability of developing DM languages that are oriented to business users.
The remainder of this paper is organized as follows. We begin with an overview of BI and DM. Next, we present the architecture and discuss its viability. We conclude presenting future research directions and conclusions.

Considerations on DM usage in BI systems

BI can be presented as an architecture, tool, technology, or system that gathers and stores data, analyses it using analytical tools, and delivers information and/or knowledge, facilitating reporting, querying, and, ultimately, allows organizations to improve decision making. A framework for BI is presented in (Azevedo

& Santos, 2009). Also relations of BI with Competitive Intelligence (CI), knowledge Management Systems (KMS), and Artificial Intelligence (AI) are explained.

The term KDD was coined in 1989 to refer to the broad process of finding knowledge in data, and to emphasize the "high-level" application of particular DM methods (Fayyad *et al.*, 1996). The DM phase concerns, mainly, to the means by which the patterns are extracted and enumerated from data.

A BI system is a particular type of system. One of the main aspects is that of user-friendly tools, that makes systems truly available to the final business user. Powerful analytical tools, such as DM, remain to complex and sophisticated for the average consumer of BI systems. (McKnight, 2002) supports that bringing DM to the front line business personnel will increase their potential to attaining BI's high potential business value. Another fundamental issue that is pointed out is the capability of DM tools to be interactive, visual, and understandable, to work directly on the data, and to be used by front line workers for intermediate and lasting business benefits.

We point out three main reasons for DM to be not completely integrated with BI. Firstly, the models/patterns obtained from DM are complex and there is the need of an analysis from a DM specialist. This fact can lead to a non-effective adoption of DM in BI, being that DM is not really integrated on most of the implemented BI systems nowadays. Secondly, the problem with DM is that, as far as our knowledge, there is not a user-friendly tool that helps decision makers to analyze DM models. Finally, but extremely important, it has not been given sufficient emphasis to the development of solutions that allow the specification of DM problems through business oriented languages, and there are also oriented for BI activities. With the expansion that has occurred in the application of DM solutions in BI this is currently of increasing importance.

Codd's relational model for database systems is long ago adopted in organizations. One of the reasons for the great success of relational databases is related with the existence of a standard language – SQL (Structured Query Language). SQL is included in all the relational database management systems (DBMS) and serves as the core above which are constructed the various graphical user interfaces (GUI's) and user friendly languages, such as Query By Example (QBE's), included in some relational DBMS.

(Imielinski & Mannila, 1996) compare file systems to the actual DM systems, and refer the aim of developing, the so called, Knowledge and Data Discovery Management Systems (KDDMS), similarly to the relational DBMS. Some efforts are being done that seek the establishment of standards in the DM area, both by academics (De Raedt, 2003; Dzeroski, 2007; Mannila, 2000) and by people in the industry field (Azevedo & Santos, 2008; Data Mining Group, 2009; Object Management Group, 2008; Tang & MacLennan, 2005; java.net, 2007). The main goal is to integrate DM with relational databases, thus allowing an easiest application of DM to business systems and making it more available to decision making.

One important issue in this domain concerns data mining languages. Several approaches have been proposed for the definition of data mining languages. The coverage of such a vast area is difficult, so we may not include it all. In the literature there can be found some language specifications, namely, DMQL (Han *et al.*, 1996), MINE RULE (Meo *et al.*, 1998), MSQL (Imielinski & Virmani, 1999), SPQL (Bonchi *et al.*, 2007), KDDML (Romei *et al.*, 2006), XDM (Meo & Psaila, 2006), RDM (De Raedt, 2002), among others. These languages are generally part of a wider system, whose aim is, undoubtedly, to achieve the already referred Knowledge and Data Discovery Management System (KDDMS), that allows the high-level abstraction existent on the relational DBMS, and that integrates the complete KDD process. Despite the importance of the referred languages, they are not business oriented. To a greater extend, they are not oriented to the diverse BI activities. This issue is of increasing importance in organizations.

An architecture for DM usage in BI systems

In this section an architecture that allows an effective usage of DM in BI systems is presented. A prototype of a BI system is implemented. The underlying relational database refers to a higher education institution and almost all the business processes of this higher education institution are supported by an information system build upon relational databases (Pereira *et al.*, 2007). At the moment the BI system delivers information about students. Examples of business questions that are put to and answered by the system are

presented on the left side of Table 1 (questions without DM). The answers to these questions are obtained by the means of OLAP tools and reporting. In a near future, DM will be integrated. This research studies the viability of implementing an architecture that conducts to an effective usage of DM in BI, allowing to bring DM to the front line business personnel and decision makers. An architecture of the BI system is presented in Figure 1.

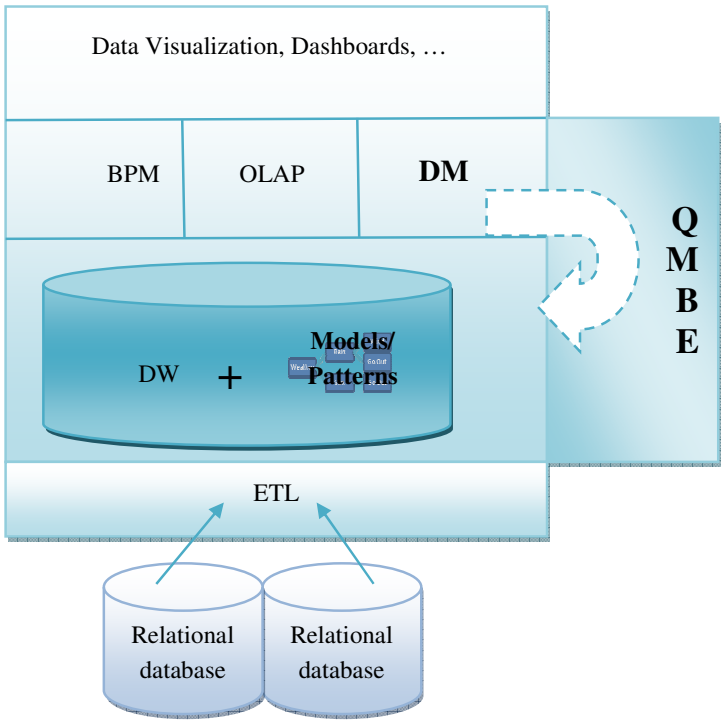

Figure 1 - Architecture for an effective usage of DM in BI

Our analysis will focus on two modules: DM module and QMBE module, since the other modules are already implemented. The DM module will feed the database with DM models. A new language, named as QMBE (Query Models By Examples) will be developed to be similar to QBE languages presented in some relational DBMS. In the right side of Table 1 examples of business questions that can be answered by the BI system after the integration of DM on the system are presented. All the questions can be converted into queries to the system defined in QBME language. The language will have two important characteristics: interactivity and iterativity.

The presented architecture consists of a conceptual architecture, which will be implemented in a near future. We consider that the system's implementation is viable because of two reasons.

The first reason is that it is according to the inductive database framework, since it will provide the features indicated in (Bonchi *et al.*, 2007), namely:

- Coupling with a DBMS, as it presents the capability for retrieving interesting data and patterns are on the same DBMS;
- Expressiveness of a query language, as it allows a high-level vision of the patterns discovery system;
- Efficiency of the mining engine;
- Graphical user interface.

Many systems have already been implemented that are according to the inductive database framework, thus providing its viability. Examples include (Bonchi *et al.,* 2007; Calders *et al.*, 2006; De Raedt, 2002; Fromont *et al.*, 2007; Han *et al.*, 1996; Imielinski & Virmani, 1999; Jamil, 2004; Kramer *et al.*, 2006; Meo & Psaila, 2006; Meo *et al.*, 1998; Romei *et al.*, 2006; Sarawagi *et al.*, 2000).

The second reason is that the actual state of the industrial standards for DM and BI will allow for the communication among the considered modules.
The system will support the complete KDD process, allowing many decisions to be made by the user in an interactive and iterative process. With the implementation of the system the authors pretend to demonstrate that the full integration of the KDD process with BI can conduct to an effective usage of DM in BI.

Questions without DM		**Questions with DM**	
Question	**Dimensions**	**Queries on models**	**Queries on models and data**
Which are the better students?	By semester By year By geographical origin By studies field …	Which are the characteristics of good students?	Select the actual students that can be good students.
Which are the worst students?		Which are the characteristics of bad students?	Select the actual students that can be bad students.
How many students conclude the grades according to initial schedule?		Which are the characteristics of the students that do not conclude the grades according to initial schedule?	Select the actual students that can not conclude the grades according to initial schedule.
Which are the courses with higher retention taxes?		Are there different types of students in the school?	
How many students exist?			

Table 1 - Examples of business questions that can be answered by the BI system

Future research directions

The presented architecture is already being implemented as a prototype. Nevertheless, there are many remaining issues for further research concerning the system's implementation. One of these, concerns the specification of the QMBE characteristics which is a critical issue. Another, concerns the definition of the communication process of the modules among each others. Finally, user interface is also a concerning.
It is expected that when tests are finished the system will be integrated in a real situation, which is that of the referred higher education institution. The implementation on a real situation will surely bring new and important inputs for improving the system.

Conclusions

We presented an architecture that allows for an effective usage of DM in BI, including a new DM language: QMBE. Our goal is to sustain that this is in fact possible. Even though the architecture was presented in a rather informal way, we believe that it can be seen as a way that conducts to DM integration with BI, thus

allowing an effective usage of DM in BI systems by business users. The interactive and iterative aspects of the system are crucial to achieve this goal.

In summary, we present a new approach to an effective usage of DM in BI systems, through the definition of a DM language that is interactive and iterative in nature. The main contribution of this paper is to present the viability of developing DM languages that are oriented to BI activities as well as oriented to business users. We hope that this preliminary research will bring fundamental issues in the field.

References

Arnott, D. & Pervan, G. (2008). Eigth Key Issues for the Decision Support Systems Discipline. Decision Support Systems, 44(3), 657-672.

Azevedo, A. & Santos, M.F. (2008). KDD, SEMMA and CRISP-DM: a Parallel Overview. In Proceedings of the IADIS European Conference on Data Mining, DM2008, 182-185.

Azevedo, A. & Santos, M.F. (2009). Business Intelligence: State of the Art, Trends, and Open Issues. In Proceedings of the First International Conference on Knowledge Management and Information Sharing - KMIS 2009, 296-300.

Bonchi, F.; Giannotti, F.; Lucchesse, C.; Orlando, S.; Perego, R. & Trasarti, R. (2007). On Interactive Pattern Mining from Relational Databases. In Dzeroski, S. & Struyf, J. (Eds.), Lecture Notes in Computer Science: Vol. 4747. Knowledge Discovery in Inductive Databases - 5th International Workshop, KDID 2006 (pp. 42-62). Berlin, Heidelberg: Springer-Verlag.

Calders, T.; Goethals, B. & Prado, A. (2006). Integrating Pattern Mining in Relational Databases. In Fürnkranz, J.; Scheffer, T. & Spiliopoulou, M. (Eds.), Lecture Notes on Artificial Intelligence: Vol. 4213. Knowledge Discovery in Databases - 10th European Conference on Principles and Practice of Knowledge Discovery in Databases - PKDD2006 (pp. 454-461). Berlin, Heidelberg: Springer-Verlag.

Clark, T. D.; Jones, M. C. & Armstrong, C.P. (2007). The Dynamic Structure of Management Support Systems: Theory Development, Research, Focus, and Direction. MIS Quarterly, 31(3), 579-615.

Data Mining Group (2009). Predictive Model Markup Language (PMML). Retrieved August, 1st 2009 from http://www.dmg.org/.

De Raedt, L. (2002). Data Mining as Constraint Logic Programming. In Kakas, A. C. & Sadri, F. (Eds.), Lecture Notes on Artificial Intelligence: Vol. 2408. Computational Logic: Logic Programming and Beyond - Essays in Honour of Robert A. kowalski - Part II (pp. 526-547). Berlin, Heidelberg: Springer-Verlag.

De Raedt, L. (2003). A perspective on Inductive Databases. SIGKDD Explorations, 4(2), 69-77.

Dzeroski, S. (2007). Towards a General Framework for Data Mining. In Dzeroski, S. & Struyf, J. (Eds.), Lecture Notes in Computer Science: Vol. 4747. Knowledge Discovery in Inductive Databases - 5th International Workshop, KDID 2006 (pp. 259-300). Berlin, Heidelberg: Springer-Verlag.

Fayyad, U. M.; Piatetski-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery: an overview. In Fayyad, U. M.; Piatetski-Shapiro, G.; Smyth, P. & Uthurusamy, R. (Eds.), Advances in knowledge discovery and data mining (pp. 1-34). Menlo Park, California: AAAI Press/The MIT Press.

Fromont, É.; Blockeel, H. & Struyf, J. (2007). Integrating Decision Tree Learning into Inductive Databases. In Dzeroski, S. & Struyf, J. (Eds.), Lecture Notes in Computer Science: Vol. 4747. Knowledge Discovery in

Inductive Databases - 5th International Workshop, KDID 2006 (pp. 81-96). Berlin, Heidelberg: Springer-Verlag.

Golfarelli, M.; Rizzi, S. & Cella, I. (2004). Beyond Data Warehousing: What`s Next in Business Intelligence. In DOLAP´04, 1-6.

Han, J.; Fu, Y.; Wang, W.; Koperski, K. & Zaiane, O. (1996). DMQL: A Data Mining Query Language for Relational Databases. In Proceedings of the SIGMOD'96 Workshop on Research Issues on Data Minining and Knowledge Discovery (DMKD'96), 27-34.

Hannula, M. & Pirttimäki, V. (2003). Business Intelligence Empirical Study on the Top 50 Finnish Companies. Journal of American Academy of Business, 2(2), 593-599.

Hoffman, T. (2009). 9 Hottest Skills for '09. Computer World, January 1(), 26-27.

Imielinski, T. & Mannila, H. (1996). A Database Perspective on Knowledge Discovery. Communications of the ACM, 39(11), 58-64.

Imielinski, T. & Virmani, A. (1999). MSQL: A Query Language for Database Mining. Data Mining and Knowledge Discovery, 3(4), 373-408.

Jamil, H. M. (2004). Declarative Data Mining Using SQL3. In Meo, R.; Lanzi, P. & Klemettinen, M. (Eds.), Lecture Notes on Artificial Intelligence: Vol. 2682. Database Support for Data Mining Applications - Discovering Knowledge with Inductive Queries (pp. 52-75). Berlin, Heidelberg: Springer-Verlag.

Kramer, S.; Aufschild, V.; Hapfelmeier, A.; Jarasch, A.; Kessler, K.; Reckow, S.; Wicker, J. & Richter, L. (2006). Inductive Databases in the Relational Model: the Data as the Bridge. In Bonchi, F. & Boulicault, J. (Eds.), Lecture Notes on Computer Science: Vol. 3933. Knowledge Discovery in Inductive Databases - 4th International Workshop - KDID2005 (pp. 124-138). Berlin, Heidekberg: Berlin-Verlag.

Mannila, H. (2000). Theoretical Frameworks for Data Mining. SIGKDD Explorations, 1(2), 30-32.

McKnight, W. (2002). Briging Data Mining to the Front Line, Part 1. Information Management magazine, November(2002), Retrieved on July, 16th 2009 at http://www.information-management.com/issues/200211001/5980-1.html.

Meo, R. & Psaila, G. (2006). An XML-Based Database for Knowledge Discovery. In Grust, T.; Höpfner, H.; Illarramendi, A.; Jablonski, S.; Mesiti, M.; Müller, S.; Patranjan, P.; Sattler; Kai-Uwe; Spiliopoulou, M. & Wijsen, J. (Eds.), Lecture Notes in Computer Science: Vol. 4254. Current Trends in Database Technology - EDTB 2006 Workshops (pp. 814-828). Berlin, Heidelberg: Springer-Verlag.

Meo, R.; Psaila, G. & Ceri, S. (1998). An Extension to SQL for Mining Association Rules. Data Mining and Knowledge Discovery, 2(2), 195-224.

Negash, S. (2004). Business Intelligence. Comunication of the Association for Information Systems, 13(), 177-195.

Object Management Group (2008). Knowledge Discovery Model (KDM). Retrieved August, 1st 2009 from http://kdmanalytics.com/kdm/index.php.

Pereira, R. H.; Azevedo, A. & Castilho, O. (2007). Secretaria On-Line From Iscap: A Case of Innovation. In Proceedings of the IADIS International Conference WWW/Internet 2007, 301-305.

Richardson, J.; Schlegel, K. & Hostmann, B. (2009). Magic Quadrant for Business Intelligence Platforms. , Gartner.

Richardson, J.; Schlegel, K.; Hostmann, B. & McMurchy, N. (2008). Magic Quadrant for Business Intelligence Platforms, 2008. , Gartner.

Romei, A.; Ruggieri, S. & Turini, F. (2006). KDDML: A Middleware Language and System for Knowledge Discovery in Databases. Data & Knowledge Engineering, 57(2), 179-220.

Sarawagi, S.; Thomas, S. & Agrawal, R. (2000). Integrating Association Rule Mining with Relational Database Systems: Alternatives and Implications. Data Mining and Knowledge Discovery, 4(2-3), 89-125.

Tang, Z. & MacLennan, J. (2005). Data Mining with SQL Server 2005. Indianapolis, IN:Wiley Publishing.

java.net (2007). jdm: Home. Retrieved August, 5th 2009 from https://jdm.dev.java.net.