

Universidade do Minho
Escola de Engenharia

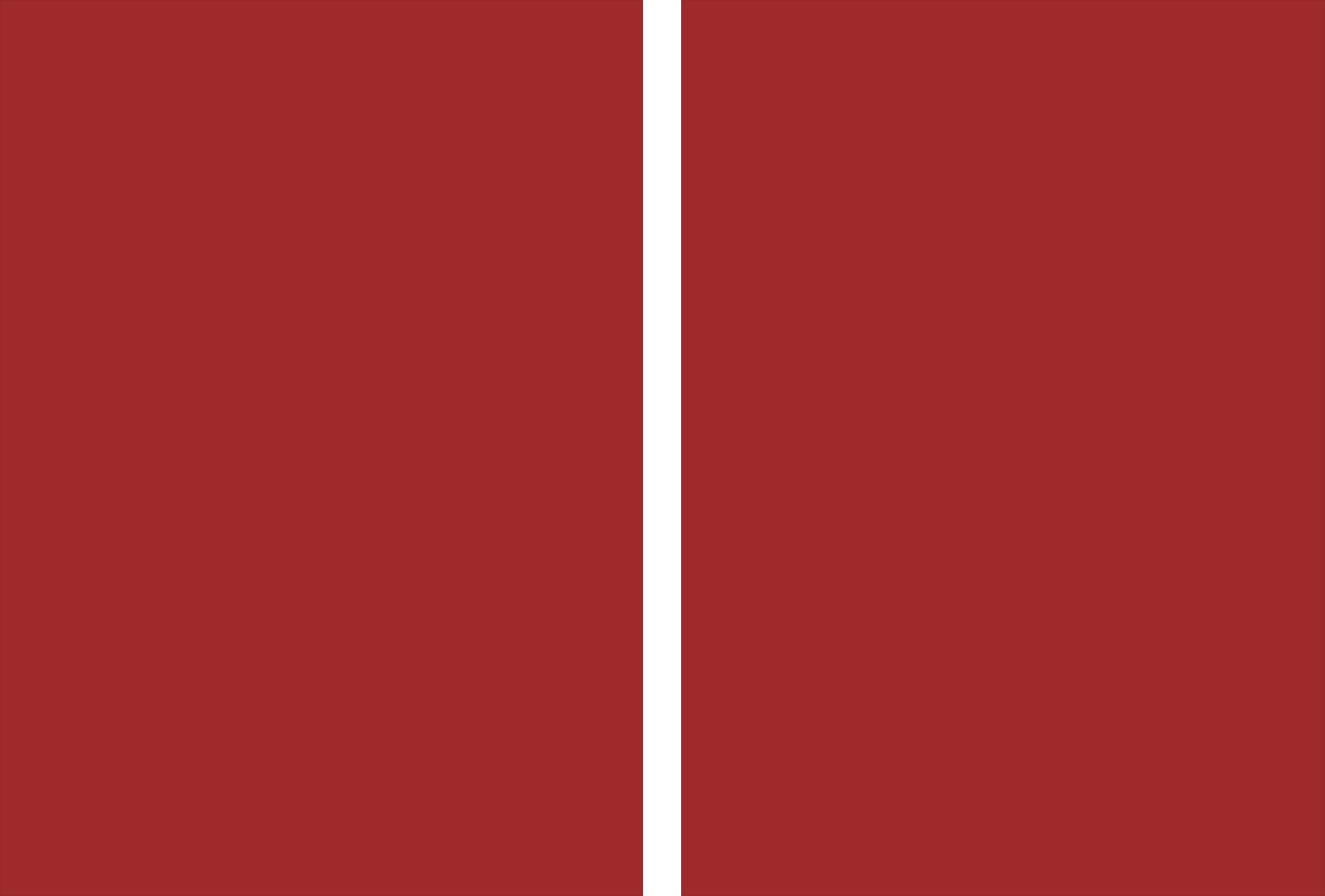
Ana Isabel Rojão Lourenço Azevedo

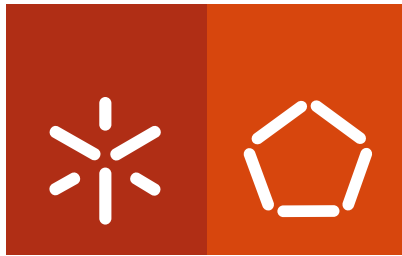
Data mining languages for business intelligence

Ana Isabel Rojão Lourenço Azevedo **Data mining languages for business intelligence**

UMinho | 2011

December 2011





Universidade do Minho
Escola de Engenharia

Ana Isabel Rojão Lourenço Azevedo

Data mining languages for business intelligence

Doctoral Thesis in Information Systems and Technologies
Area of Engineering and Management Information Systems

Supervised by
Professor Manuel Filipe Santos

December 2011

É AUTORIZADA A REPRODUÇÃO INTEGRAL DESTA TESE APENAS PARA EFEITOS DE INVESTIGAÇÃO, MEDIANTE DECLARAÇÃO ESCRITA DO INTERESSADO, QUE A TAL SE COMPROMETE;

Universidade do Minho, ___/___/_____

Assinatura: _____

Acknowledgments

First of all, to my family. They were a fundamental piece during the development of this project. I missed the moments that could be spent with them. I am thankful for their understanding and collaboration. A special word goes to my husband, for being my bridge over troubled waters during the difficult times. Also to my three children, who are my inducement. To my parents, who have always been my biggest fans, without them none of this would be possible.

To my supervisor who helped me along the way. Thank you for the ideas, the discussions, and the countless patience.

To Professor João Álvaro Carvalho who kindly welcomed me at DSI.

To my friends and colleagues who helped in a way or another. I specially render thanks to António Abreu and to Rosalina Babo whose collaboration was of inestimable value.

To the Information Systems division of ISCAP who helped me with the database and the involved issues, namely to Rui Pereira, Tiago Moreira, and Vitor Silva.

To ISCAP's administration who authorized me to access the necessary data for the development of this project.

.

RESUMO

Desde que Lunh usou, pela primeira vez, em 1958, o termo Business Intelligence (BI), grandes transformações se operaram na área dos sistemas e tecnologias de informação e, em especial, na área dos sistemas de apoio à decisão. Atualmente, os sistemas de BI são amplamente utilizados nas organizações e a sua importância estratégica é largamente reconhecida. Estes sistemas apresentam-se como essenciais para um completo conhecimento do negócio e como uma ferramenta insubstituível no apoio à tomada de decisão. A divulgação das ferramentas de Data Mining (DM) tem vindo a aumentar na área do BI, assim como o reconhecimento da relevância da sua utilização nos sistemas de BI empresariais.

As ferramentas de BI são ferramentas amigáveis, iterativas e interativas, permitindo aos utilizadores finais um acesso fácil. Desta forma, é possível ao utilizador final manipular diretamente os dados, tendo assim a possibilidade de extrair todo o valor para o negócio neles contido. Um dos problemas apontados na utilização do DM na área do BI prende-se com o facto de os modelos de DM serem, em geral, demasiado complexos para que os utilizadores de negócio os possam manipular diretamente, contrariamente ao que ocorre com as outras ferramentas de BI.

Neste contexto, foi identificado como problema de investigação a não existência de ferramentas de BI que possibilitem ao utilizador de negócio a manipulação direta dos modelos de DM e, conseqüentemente, não possibilitando extrair todo o valor potencial neles contidos. Este aspeto reveste-se de particular importância num universo empresarial no qual a concorrência é cada vez mais forte e no qual o conhecimento do negócio, das variáveis envolvidas e dos potenciais cenários representam um papel fundamental para as organizações poderem concorrer num mercado extremamente exigente.

Considerando que os sistemas de BI assentam, maioritariamente, sobre sistemas operacionais que utilizam sobretudo o modelo relacional de bases de dados, a investigação efetuada inspirou-se nos conceitos ligados ao modelo relacional de bases de dados e nas linguagens a ele associadas em particular as linguagens Query-By-Example (QBE). Estas linguagens têm uma forte componente de interactividade, são amigáveis e permitem iteratividade e são amplamente utilizadas em ambiente de negócio pelos utilizadores finais.

Têm vindo a ser desenvolvidos esforços no sentido do desenvolvimento de padrões e normas na área do DM, sendo dada grande relevância ao tema das bases de dados indutivas. No contexto

das bases de dados indutivas é dada grande relevância às chamadas linguagens de DM. Estes conceitos serviram, igualmente, de inspiração a esta investigação. Apesar da importância destas linguagens de DM, elas não estão orientadas para os utilizadores finais em ambientes de negócio.

Ligando os conceitos relacionados com as linguagens QBE e com as linguagens de DM, foi concebida e implementada uma linguagem de DM para BI, à qual foi dado o nome QMBE. Esta nova linguagem é por natureza amigável, iterativa e interativa, isto é, apresenta as mesmas características que as ferramentas de BI habituais permitindo aos utilizadores finais a manipulação direta dos modelos de DM e, deste modo, aceder a todo o valor potencial desses modelos com todas as vantagens que daí poderão advir. Utilizando um protótipo de um sistema de BI, a linguagem foi implementada, testada e avaliada conceptualmente. Verificou-se que a linguagem possui as propriedades desejadas, a saber, é amigável, iterativa, interativa. Finalmente, a linguagem foi avaliada por utilizadores finais que já tinham experiência anterior na utilização de DM em contexto de BI. Verificou-se que na ótica destes utilizadores a utilização da linguagem apresenta vantagens em relação à utilização tradicional de DM no âmbito do BI.

Palavras-chave: Business Intelligence, Descoberta de Conhecimento em Bases de Dados, Data Mining, Linguagens de Data Mining, Query-By-Example, Bases de Dados Indutivas, Data Warehouses Indutivas, Modelo Relacional, Design Science Research.

ABSTRACT

Since Lunh first used the term Business Intelligence (BI) in 1958, major transformations happened in the field of information systems and technologies, especially in the area of decision support systems. Nowadays, BI systems are widely used in organizations and their strategic importance is clearly recognized. These systems present themselves as an essential part of a complete knowledge of business and an irreplaceable tool in the support to decision making. The dissemination of data mining (DM) tools is increasing in the BI field, as well as the acknowledgement of the relevance of its usage in enterprise BI systems.

BI tools are friendly, iterative and interactive, allowing business users an easy access. This way, the user can directly manipulate data, thus having the possibility to extract all the value contained into that business data. One of the problems noted in the use of DM in the field of BI is related to the fact that DM models are, generally, too complex in order to be directly manipulated by business users, as opposite to other BI tools.

Within this context, the nonexistence of BI tools allowing business users the direct manipulation of DM models was identified as the research problem, since that, as a consequence of business users not directly manipulating DM models, they can be not able of extracting all the potential value contained in DM models. This aspect has a particular relevance in an entrepreneurial universe where competition is stronger every day and the knowledge of the business, the variables involved and the possible scenarios play a fundamental role in allowing organizations to compete in an extremely demanding market.

Considering that the majority of BI systems are built on top of operational systems, which use mainly the relational model for databases, the research was inspired on the concepts related to this model and associated languages in particular Query-By-Example (QBE) languages. These languages are widely used by business users in business environments, and have got a strong interactivity component, are user-friendly, and allow for iterativeness.

Efforts are being developed in order to create standards and rules in the field of DM with great relevance being given to the subject of inductive databases. Within the context of inductive databases a great relevance is given to the so called DM languages. These concepts were also an inspiration for this research. Despite their importance, these languages are not oriented to business users in business environments.

Linking concepts related with QBE languages and with DM languages, a new DM language for BI, named as Query-Models-By-Example (QMBE) was conceived and implemented. This new language is, by nature, user-friendly, iterative and interactive; it presents the same characteristics as the usual BI tools allowing business users the direct manipulation of DM models and, through this, the access to the potential value of these models with all the advantages that may arise. Using a BI system prototype, the language was implemented, tested, and conceptually evaluated. It has been verified that the language possesses the desired properties, namely, being user-friendly, iterative, and interactive. The language was evaluated later by business users who were already experienced in using DM within the context of BI. It has been verified that, according to these users, using the language presents advantages when comparing to the traditional use of DM within BI.

Keywords: Business Intelligence, Knowledge Discovery from Databases, Data Mining, Data Mining Standards, Data Mining Languages, Query-By-Example, Inductive Databases, Inductive Data Warehouses, Relational Model, Business Users, Design Science Research.

Table of Contents

Acknowledgments.....	i
RESUMO	iii
ABSTRACT	v
Table of Contents	vii
Terminology	xi
List of Figures.....	xiii
List of Tables.....	xv
PART I – Presentation	1
1 Introduction	3
1.1 Motivation.....	3
1.2 Research Objectives and Contribution	6
1.3 Thesis Organization	9
Part II – Background.....	11
2 Business Intelligence.....	13
2.1 Business Intelligence Roots and Associations	13
2.2 Research on Business Intelligence	15
2.3 A Framework for Business Intelligence	17
2.4 An Example of a Business Intelligence System.....	18
3 Data Mining and Knowledge Discovery in Databases.....	21
3.1 The Knowledge Discovery in Databases Process.....	21
3.2 Some Applications	22
3.3 Data Mining Tasks, Methods/Algorithms, and Models/Patterns	25
4 Data Mining Languages.....	31
4.1 Towards Standards for Data Mining.....	31
4.1.1 Industrial Standards	31
4.1.2 Scientific Research	35
4.2 Inductive Databases and Data Mining Languages	36

4.2.1	Special Purpose Languages.....	37
4.2.2	Languages Using Standard SQL.....	40
4.3	Data Mining Integration with Relational Databases.....	40
5	Query-By-Example Languages.....	42
5.1	General Notions.....	43
5.2	Relational Calculus and Query-By-Example Languages.....	48
Part III – Research Approach and Research Outputs.....		51
6	Research Problem.....	53
7	Research Framework.....	56
8	Research Methodology.....	59
9	Inductive Data Warehouse.....	64
9.1	An example of an inductive Data Warehouse.....	64
9.2	Generalization.....	67
10	Query-Models-By-Example Language.....	70
10.1	Queries on data.....	71
10.1.1	Using an example.....	71
10.1.2	A general query.....	74
10.2	Queries on models.....	75
10.2.1	Using an example.....	75
10.2.2	A general query.....	77
10.3	Queries on models and data.....	78
10.3.1	Using an example.....	78
10.3.2	A general query.....	79
10.4	Relational calculus and QMBE.....	80
10.4.1	Traditional QBE: Queries on data.....	80
10.4.2	QMBE Extension 1: queries on models.....	81
10.4.3	QMBE Extension 2: queries on models and data.....	82
11	Query-Models-By-Example Evaluation.....	83
11.1	Conceptual evaluation.....	83
11.2	Questionnaire to business users.....	84
11.2.1	Questionnaire structure.....	84
11.2.2	Analysis of the questionnaire results.....	86
11.3	Some brief considerations.....	94

Part IV – Final Conclusions	95
12 Discussion and Related Work	97
13 Conclusion and Future Research Directions	99
13.1 Thesis Contributions	99
13.2 Critical reflection about the obtained results and future work	101
References.....	103
Appendix A – Questionnaire	119
Appendix B – Statistical Tests	123
Appendix C – List of published papers.....	127
Appendix D –RIPPER Algorithm.....	129
Appendix E – The data mining model obtained using RIPPER	131

Terminology

ADS	Automatic Decision System
AI	Artificial Intelligence
BAM	Business Activity Management
BI	Business Intelligence
BPM	Business Performance Management
CI	Competitive Intelligence
CRISP-DM	Cross-Industry Standard Process for Data Mining
DB	Database(s)
DBMS	Database Management System
DM	Data Mining
DSR	Design Science research
DSS	Decision Support System(s)
DW	Data Warehouse
EIS	Executive Information Systems
ETL	Extract Transform and Load
GIS	Geographical Information System
GUI	Graphical User Interface
HEI	Higher Education Institution
IDB	Inductive Database
IDW	Inductive Data Warehouse
IS	Information System
KB	Knowledge Base
KDD	Knowledge Discovery on Databases
KDDMS	Knowledge and Data Discovery Management System
KMS	Knowledge Management System
MIS	Management Information System
OLAM	On-Line Analytical Mining
OLAP	On-Line Analytical Processing
QBE	Query-By-Example
QMBE	Query-Models-By-Example

RDBMS	Relational Database Management Systems
SEMMA	Sample, Explore, Modify, Model, Assess
SOA	Service Oriented Architecture
SQL	Structured Query Language
XML	Extensible Markup Language

List of Figures

Figure 1 – Database Languages, DM languages, DM languages for BI	5
Figure 2 – Inductive Data Warehouse.....	7
Figure 3 – BI Associations.....	15
Figure 4 – A Framework for Business Intelligence.....	18
Figure 5 – High-level architecture of a Business Intelligence System	19
Figure 6 – The KDD Process.....	22
Figure 7 - The CRISP-DM life cycle (Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer & Wirth, 2000 - pp 13) 33	33
Figure 8 – Relational Database Schema	43
Figure 9 – Skeleton table for a QBE language.....	44
Figure 10 – QBE query 1	45
Figure 11 – QBE query 1 result/answer.....	45
Figure 12 – QBE query 2	45
Figure 13 – QBE query 2 result/answer.....	46
Figure 14 – QBE query 3	46
Figure 15 – QBE query 3 result/answer.....	47
Figure 16 – QBE query 4	47
Figure 17 – QBE query 4 result/answer.....	48
Figure 18 – Relational calculus propositions for Q1 and Q2.....	49
Figure 19 – Information Systems (IS) research framework (Hevner, 2004 – pp 80)	51
Figure 20 – Research Problem.....	55
Figure 21 – Research Framework	57
Figure 22 – Architecture for integration of Data Mining with Business Intelligence	60
Figure 23 – DW schema used in the research	62
Figure 24 – Model Table schema used in the research.....	66
Figure 25 – IDW schema used in the research	66
Figure 26 – Examples of rules and their corresponding representation in the Model Table.....	67
Figure 27 – IDW General Schema	68
Figure 28 – Skeleton table for the QMBE language.....	70
Figure 29 – Types of QMBE queries	71
Figure 30 – QMBE Query 1.....	72
Figure 31 – Answer to query 1	72
Figure 32 – QMBE Query 2.....	72
Figure 33 – Answer to query 2	73
Figure 34 – QMBE Query 3.....	73
Figure 35 – Answer to query 3	73
Figure 36 – QMBE Query 4.....	74
Figure 37 – Answer to query 4	74
Figure 38 – QMBE query I	75
Figure 39 – QMBE Query 5.....	76
Figure 40 – Answer to query 5.....	76
Figure 41 – QMBE Query 6.....	76
Figure 42 – Answer to query 6.....	76
Figure 43 – QMBE Query 7.....	77
Figure 44 – Answer to query 7	77
Figure 45 – QMBE query J.....	77
Figure 46 – QMBE Query 8.....	79
Figure 47 – Answer to query 8.....	79
Figure 48 – QMBE query K.....	80
Figure 49 – Respondents’ experience about the importance of DM (Graph)	87
Figure 50 – Respondents’ opinion about adopting QMBE in their organizations (Graph)	93
Figure 51 – Research contributions.....	101

List of Tables

Table 1 – Examples of business questions that can be answered by the Business Intelligence System	20
Table 2 – Outline of DM tasks, Methods/Algorithms, Models/Patterns, and Guidance.....	30
Table 3 – Summary of the correspondences between KDD, SEMMA and CRISP-DM.....	34
Table 4 – Comparison of SQL-Based languages syntax	39
Table 5 – Examples of business questions involving DM models	61
Table 6 – Questionnaire goals and associated questions	86
Table 7 – Respondents' experience using BI and using DM	87
Table 8 – Respondents' general reaction to QMBE	88
Table 9 – Comparison of respondents' opinions about using DM vs using DM with QMBE.....	92

PART I – PRESENTATION

This thesis presents the key issues of the research developed in the ambit of the Doctoral Program in Information Systems and Technologies, held in the University of Minho. It begins with the introduction, in chapter 1. This chapter starts with the presentation of the motivation that lead to the research, in Section 1.1, continues with the introduction of the research objectives and contributions intended to be achieved, in Section 1.2, and with the thesis organization, in Section 1.3.

1 Introduction

The project presented in this thesis approaches the issue of using Data Mining (DM) languages in the context of Business Intelligence (BI) systems. The aim is to study the viability of developing a DM language oriented to business users and oriented to the BI activities.

1.1 Motivation

Business Intelligence (BI) is one emergent area of the Decision Support Systems (DSS) discipline and can be defined as the process that transforms data into information and then into knowledge (Golfarelli, Rizzi & Cella, 2004). Being rooted in the DSS discipline, BI has suffered a considerable evolution over the last 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, Jones & Armstrong, 2007; Davenport, 2010; Hannula & Pirttimäki, 2003; Hoffman, 2009; Negash, 2004; Richardson, Schlegel & Hostmann, 2009; Richardson, Schlegel, Hostmann & McMurchy, 2008; Sallam, Hostman, Richardson & Bitterer, 2010). 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.

The term Knowledge Discovery in Databases (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 data mining (DM) methods (Fayyad, Piatetski-Shapiro & Smyth, 1996). The DM phase concerns, mainly, to the means by which patterns are extracted and enumerated from data.

DM is being applied with success in BI and several examples of applications can be found (Linoff, 2008; Turban, Sharda, Arosen & King, 2008; Vercellis, 2009). Despite that, DM has not yet reached to non specialized users and thus it is not yet completely integrated with BI. Powerful analytical tools, such as DM, remain too complex and sophisticated for the average consumer of BI systems. McKnight supports that bringing DM to the front line business personnel will increase their potential to attaining BI’s high potential business value (McKnight, 2002). Another fundamental issue that is pointed out by McKnight 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.

Currently, DM systems are functioning as separate isles, and hereby it is considered that only the full integration of the KDD process on BI can conduct to an effective usage of DM in BI (Azevedo & Santos, 2011). Three main reasons can be pointed out for DM to be not completely integrated with BI, each one leading to a specific problem that constraints DM usage in 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 there is not a user-friendly tool that can be used by decision makers to analyze DM models. Usually, BI systems have user-friendly analytical tools that help decision makers in order to obtain insights on the available data and allow them to take better decisions. Examples of such tools are On-Line Analytical Processing (OLAP) tools, which are widely used (Negash, 2004; Turban, Sharda, Arosan & King, 2008). There are not equivalent tools for DM that allow business users to obtain insights in 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 that 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.

BI systems are, usually, built on top of relational databases and diverse types of languages are involved (Figure 1). As a consequence, DM integration with relational databases is an important issue to consider when studying DM integration with BI. Codd 's relational model for database systems (Codd, 1970; Codd, 1982) has been adopted long ago in organizations. One of the reasons for the great success of relational databases is related with the existence of a standard language – Structured Query Language (SQL). SQL allows business users to obtain quick answers to ad-hoc business questions, through queries on the data stored in databases. SQL is nowadays included in all the Relational Database Management Systems (RDBMS). SQL serves as the core above which are constructed the various Graphical User Interfaces (GUI) and user friendly languages, such as Query-By-Example (QBE), included in RDBMS (Date, 2004; Elmasri & Navathe, 2007). It is also necessary to define a standard language, which can operate likewise for data mining. Several approaches have been proposed for the definition of data mining languages. In the literature there can be found some language specifications, namely, DMQL (Han, Fu, Wang, Koperski & Zaiane, 1996), MINE RULE (Meo, Psaila & Ceri, 1998), MSQL (Imielinski & Virmani, 1999), SPQL (Bonchi, Giannotti, Lucchesse, Orlando, Perego &

Trasarti, 2007), KDDML (Romei, Ruggieri & Turini, 2006), XDM (Meo & Psaila, 2006), RDM (De Raedt, 2002), among others. 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 nowadays.

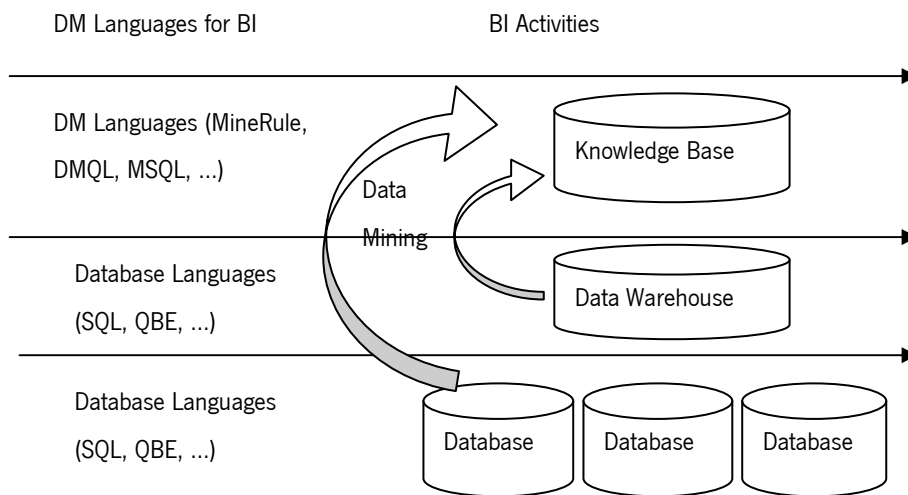


Figure 1 – Database Languages, DM languages, DM languages for BI

DM integration with BI systems can be tackled from different perspectives. On the one hand, it can be considered that the effective integration of DM with BI systems must involve final business users' access to DM models. This access is crucial in order to business users to develop an understanding of the models, to help them in decision making. Han and Kamber state that the integration (coupling) of DM with database systems and/or data warehouses is crucial in the design of DM systems (Han & Kamber, 2006). They consider four possible integration schemes, which are, in increasing order of integration: no coupling, louse coupling, semi-tight coupling, and tight coupling. They present the concept of On-Line Analytical Mining (OLAM), which incorporates OLAP with DM, as a way to achieve tight coupling. On the other hand, a different approach can be considered, through the outgrowth of new strategies that allow business users and DM specialists developing new communication strategies. Wang and Wang introduce a model that allows knowledge sharing among business insiders and DM specialists (Wang & Wang, 2008). It is argued that this model can make DM more relevant to BI.

The developed research, presented in this thesis, focus on making DM models to be directly manipulated by business users. It is considered that this can conduct to an understanding of DM models by business users, helping them on the decision making process. The fact that DM models can be directly manipulated by business users, can help unblocking the potential business value hidden in DM models. With this in mind, a high-level architecture that intends to conduct to an effective usage of DM with BI is presented (Azevedo & Santos, 2009b). This architecture includes a new DM language, named as Query-Models-By-Example (QMBE) that is iterative and interactive in nature, thus allowing business users to directly access and manipulate DM models. In this research, it is considered that allowing business users to directly manipulate DM models is a fundamental issue. A great amount of business value is hidden in DM models. Sometimes this business value can be not discovered during the analysis of DM specialists, since they, usually, have an insufficient knowledge of the involved business issues. On the other hand, this value can be discovered if business users can be able to directly access DM models. This access is considered the privileged way that allows business users accessing and exploring all the potential value of DM models. This will surely bring advantages to the process of decision making in organizations.

1.2 Research Objectives and Contribution

BI is the top level of a complex system (Figure 1). On its foundations lay several databases, usually based in the relational model (Codd, 1970) for databases (DB), that can be accessed and manipulated using specific database (DB) languages, such as SQL and QBE. On the next level, data warehouses (DW) can be manipulated using exactly the same sort of languages. Applying DM to data stored on both DB and DW¹, knowledge bases (KB) arise on the next level. KB store DM models and, traditionally, are not based on the relational model, unlike DB and DW. Nevertheless, using the framework of inductive databases (IDB), DM models can be stored in databases in the same way as data, thus DM models can be accessed and manipulated at the same level than data (De Raedt, 2003; Dzeroski, 2007; Imielinski & Mannila, 1996). Using the framework of IDB, DM models can be obtained and manipulated through the use of DM languages, such as MineRule (Meo, Psaila & Ceri, 1998), DMQL (Han, Fu, Wang, Koperski & Zaiane, 1996), or MSQL (Imielinski & Virmani, 1999). Despite the importance of these

¹ Being aware that a data warehouse is, first of all, a database, the term **database** is here used as synonym of **operational database**.

languages, they are not business oriented, are not oriented to business users and are not oriented to BI activities. This is a crucial issue in organizations that is gaining momentum each day.

In the context of BI there can be said that an IDB contains both the DW and the KB (Figure 2) and thus we can refer to this database as Inductive Data Warehouse (IDW). This is an important concept in the realm of this research since it focuses on making DM models available to business users. In an IDW data and DM models are stored at the same level and thus DM models can be accessed by business users in the same way as data.

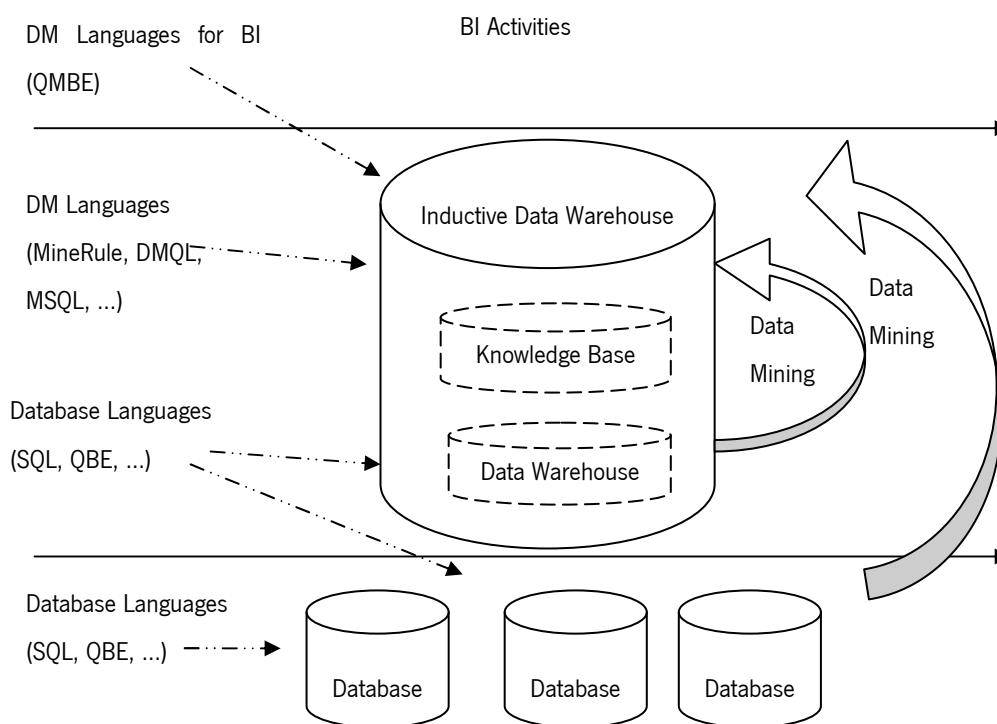


Figure 2 – Inductive Data Warehouse

The importance of allowing final business users to access and manipulate DM models comes up from the need of allowing business users to be more autonomous, without the permanent necessity to depend on the presence of a DM specialist. Moreover, considering that DM specialists do not usually have a complete knowledge of the business issues, making DM directly available to business users is the key element that allows obtaining all the potential business value hidden in DM models. This can be done by means of a DM language developed, above all, to accomplish the necessities of final business users of BI systems.

Consequently, it is considered in the research hereby presented the importance of developing DM languages for BI, which are oriented to business users and, moreover, to BI activities.

Considering the above mentioned aspects, the study presented in this thesis, aims to analyze the viability of developing a DM language oriented to BI activities as well as oriented to business users. The developed DM language is designed so that it can be used by business users in order to directly manipulate DM models, thus being able to explore their potential value. It is also intended to study possible impacts in the bottom levels of the system and thus define requisites in each one of the presented levels.

Despite that a deeper presentation of the research contributions is developed in chapter 13, some of them are now introduced. The main contribution of the research presented in this thesis is to allow business users to directly manipulate DM models, thus being able to completely explore their potential business value. This is achieved by means of the use of the IDB framework in the area of BI, presenting the concept of IDW as a DW storing both data and DM models at the same level. The use of the IDB framework in the area of BI can be considered a novelty. One important point in IDB is the DM language. The same goes for IDW. Hereby it is presented a new data mining language for BI, named as Query-Models-By-Example (QMBE), which is oriented to BI activities as well as oriented to business users. QMBE is presented as an extension of traditional QBE languages, which are included in most of the RDBMS nowadays. As a consequence, this new DM language is iterative and interactive in nature. It allows business users to answer to ad-hoc business questions through queries on data or/and on DM models. QMBE allows business users to directly access and manipulate DM models. The novelty of the QMBE language is that it is oriented to business users and to BI activities. This kind of approach allows business users to directly access and manipulate data and models. This will bring DM to the business users, alike other BI tools, allowing them to completely exploring DM potential value. The research presented in this thesis was developed accordingly to the principles of design science research (DSR). Therefore, some insights are provided on DSR usage in the Information Systems (IS) discipline. DSR is a new trend in the IS discipline, thus it is important to reflect about it and to develop research based on its principles. Consequently the research presented in this thesis can be considered as a contributions to a better understanding of the application of DSR in the IS discipline.

1.3 Thesis Organization

The remaining of the thesis is organized as follows. In part II, as a result from the literature review, some general concepts related to the thesis fields of study are presented, namely, Business Intelligence, in chapter 2, Data mining and Knowledge Discovery in Databases, in chapter 3, Data Mining Languages, in chapter 4, and Query-By-Example languages, in chapter 5. In part III, the research approach and the research outputs are described, starting with the research problem, in chapter 6, presenting next the research framework, in chapter 7, and describing the research methodology, in chapter 8. Two important research outputs are then introduced, namely the concept of inductive data warehouse, in chapter 9, and a proposal for a new data mining language named as QMBE, in chapter 10. Part III ends with QMBE evaluation, in chapter 11. The thesis ends with part IV, which contains discussion and related work, in chapter 12, and closes with conclusion and future research directions, in chapter 13.

PART II – BACKGROUND

As a result of the literature review, some general concepts related to the thesis fields of study are introduced. It is initiated with an overview of the area of business intelligence, in chapter 2, presenting business intelligence roots and associations, in Section 2.1, an overview of research on business intelligence, in Section 2.2, a framework for business intelligence, in Section 2.3, and an example of a business intelligence system, in Section 2.4. It proceeds, in chapter 3, with some concepts related with data mining and the knowledge discovery in databases process by, first, clarifying notions, in Section 3.1, next, exploring some applications, in Section 3.2, and ending by making a summary of Data Mining (DM) tasks, methods/algorithms, and models/patterns, in Section 3.3. Next, in chapter 4, DM languages are introduced, being that standards for DM, in Section 4.1, are the point of departure that leads to inductive databases and data mining languages, in Section 4.2, and to DM integration with relational databases, in Section 4.3. Chapter 5 concludes Part I presenting Query-By-Example (QBE) languages. It starts with the necessary general notions, in Section 5.1, and then relating QBE languages with relational calculus, in Section 5.2.

2 Business Intelligence

Business Intelligence (BI) can be presented as an architecture, a tool, a technology or a system that gathers and stores data, analyzes it using analytical tools, and delivers information and/or knowledge, facilitating reporting, querying and, ultimately, allowing organizations to improve decision making (Clark, Jones & Armstrong, 2007; Kudyba & Hoptroff, 2001; Michalewicz, Schmidt, Michalewicz & Chiriac, 2007; Moss & Shaku, 2003; Negash, 2004; Raisinghani, 2004; Steiger, 2010; Thierauf, 2001; Turban, Sharda, Arosen & King, 2008). To put it shortly, Business Intelligence (BI) can be defined as the process that transforms data into information and then into knowledge (Golfarelli, Rizzi & Cella, 2004). More recently, in (Michalewicz, Schmidt, Michalewicz & Chiriac, 2007) the notion of Adaptive Business Intelligence is presented, incorporating Artificial Intelligence (AI) with BI.

Being rooted in the Decision Support Systems (DSS) discipline, BI has suffered a considerable evolution over the last 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, Jones & Armstrong, 2007; Hannula & Pirttimäki, 2003; Hoffman, 2009; Negash, 2004; Richardson, Schlegel & Hostmann, 2009; Richardson, Schlegel, Hostmann & McMurchy, 2008; Sallam, Hostman, Richardson & Bitterer, 2010). BI has strong associations with Knowledge Management (KM) and Competitive Intelligence (CI) (Clark, Jones & Armstrong, 2007; Liebowitz, 2006; Negash, 2004; Turban, Sharda, Arosen & King, 2008; Zeller, 2008). Despite being treated as independent areas, the intersections between them must be considered.

2.1 Business Intelligence Roots and Associations

The roots for Business Intelligence (BI) can be found in the field of Decision Support Systems (DSS), which “is the area of the information systems (IS) discipline that is focused on supporting and improving managerial decision-making” (Arnott & Pervan, 2008). DSS can also be presented as a computer-based solution that can be used to support complex decision making, and solving complex, semi-structured, or ill-structured problems (Nemati, Steiger, Iyer & Herschel, 2002; Shim, Warkentin, Courtney, Power, Sharda & Carlsson, 2002). The term BI has replaced other terms such as Executive Information Systems (EIS) and Management Information Systems (MIS) (Negash, 2004; Turban, Sharda, Arosen & King, 2008). Nowadays it is possible to say that BI is an area of DSS that attracts a great deal of interest. BI refers to

Information Systems aimed at integrating structured and unstructured data in order to convert it into useful information and knowledge, upon which business managers can make more informed and consequently better decisions.

BI is associated with Competitive Intelligence (CI) and Knowledge Management (KM) systems (Clark, Jones & Armstrong, 2007; March & Hevner, 2007; Negash, 2004; Thierauf, 2001; Turban, Sharda, Arosan & King, 2008). Negash presents CI as a branch of BI, and refers to it as “a systematic and ethical program for gathering, analyzing and managing external information that can affect company’s plans, decisions and operations” (Negash, 2004 - pp 186). KM systems refer to “IT-based systems developed to support and enhance the organizational processes of knowledge creation, storage/retrieval, transfer, and application.” (Alavi & Leidner, 2001 – pp. 114). It can be argued that BI and KM systems are not disparate systems, but that they are complementary as they share elements required to support managerial decision making (Clark, Jones & Armstrong, 2007; Liebowitz, 2006). It can also be argued that “KM and BI, while differing, need to be considered together as necessarily integrated and mutually critical components in the management of intellectual capital” (Herschel & Jones, 2005 – pp 45). Moreover, BI, KM, CI, and AI should be aggregated so as “to provide value-added information and knowledge toward making organizational strategic decisions” (Liebowitz, 2006 - pp 22), in order to achieve Strategic Intelligence for businesses (Figure 3).

“Organizational performance often depends more on an ability to turn knowledge into effective action and less on knowledge itself” (Alavi & Leidner, 2001 – pp 129). Deeper studies involving the associations presented could lead to an understanding of how BI could lead decision makers to attain this ability.

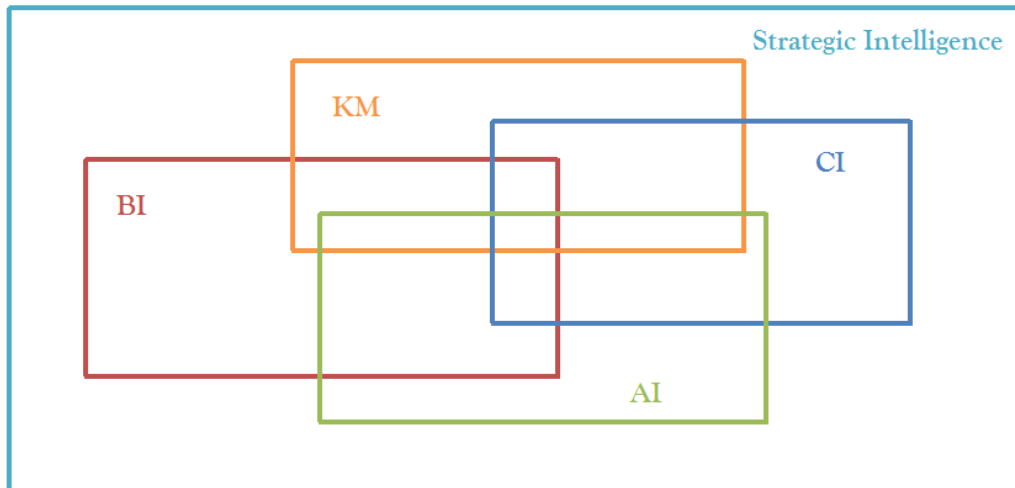


Figure 3 – BI Associations

2.2 Research on Business Intelligence

Despite the wide acceptance that the term BI was coined by Gartner in 1989 (Power, 2007; Turban, Arosan, Liang & Sharda, 2007; Turban, Sharda, Arosan & King, 2008; Zeller, 2007), the first reference to Business Intelligence was made by Lunh (Lunh, 1958), and several publications on BI can be found between 1958 and 1989. Lately, the use of the term BI has been growing (Arnott & Pervan, 2008; Hannula & Pirttimäki, 2003), and there can be found a significant number of publications that focus on this subject, as well as professional associations whose main goal is to disseminate the use of BI throughout organizations. Software vendors have defined positions on the market with diversified BI software packages and open source platforms are also available. As a result, the market tends to stabilize (Richardson, Schlegel & Hostmann, 2009; Richardson, Schlegel, Hostmann & McMurchy, 2008; Sallam, Hostman, Richardson & Bitterer, 2010).

Negash, in 2004, refers that Information Systems research in the BI field was, by that time, scarce (Negash, 2004). Since then, scientific research is growing at a significant rate, as can be confirmed by a search in some of the most popular scientific sources. A great number of publications about BI appear in diversified publications and new journals focused specifically on BI are arising. The literature presents research that explores several aspects of BI. Following, a few will be mentioned:

- In (Hannula & Pirttimäki, 2003) an empirical study about BI activities in Finnish Companies is presented;

- In (Arnott & Pervan, 2008; Clark, Jones & Armstrong, 2007), and (March & Hevner, 2007) the authors include in their research references to the role of BI in the DSS and IS disciplines;
- In (Pervan & Arnott, 2006) an analysis is made on research in data warehousing and BI between 1990 and 2004;
- In (Cheng, Lu & Sheu, 2009), and (Li, Shue & Lee, 2008) the authors develop BI applications to specific managerial problems;
- In (Elbashir, Collier & Davern, 2008), and (Lin, Tsai, Shiang, kuo & Tsai, 2009) the authors intend to develop models to evaluate BI systems;
- In (Hobek, Ariyachandra & Frolick, 2009), and (Watson, 2009) the authors are concerned about the role that people play on a BI project.

It is difficult to be comprehensive on the coverage of such a vast area hence a choice was made to highlight the trends and research issues considered most relevant. One trend is Pervasive BI, or BI for the masses (Eckerson, 2008; Lunger, 2008; Negash, 2004). There is a concern on delivering BI to all levels of an organization. Another trend is Real-time BI or Operational BI, which intends to deliver information based on real time data, as opposed to historical data (Brobst & Pareek, 2009; Klawans, 2008; Negash, 2004). Other point concerns on how to deal with the increasing quantities of data available for BI systems (Klawans, 2008; Strenger, 2008). Emphasis is also being placed on cultural aspects and on the human side of BI (Hobek, Ariyachandra & Frolick, 2009; Lin, Tsai, Shiang, kuo & Tsai, 2009; Watson, 2009). Some research issues that have been identified in the literature on DSS could also be explored in the BI area, namely, integration issues, analysis of usability, assessment, return on investment, and technological issues. A research area could analyze and evaluate technologies that are potentially applicable to BI analysis and understanding (Nemati, Steiger, Iyer & Herschel, 2002). Powerful analytical tools, such as DM, remain too complex and sophisticated for the average consumer, therefore, another area of research could be the development of more effective human-computer interfaces (Azevedo & Santos, 2009a; Clark, Jones & Armstrong, 2007).

2.3 A Framework for Business Intelligence

As pointed out above, BI refers to information systems aimed at integrating structured and unstructured data in order to convert it into useful information and knowledge, upon which business managers can make more informed and consequently better decisions. There are different approaches to BI:

- The traditional approach to BI is concerned with data aggregation, business analytics and data visualization (Kudyba & Hoptroff, 2001; Raisinghani, 2004; Turban, Sharda, Aroson & King, 2008). According to this approach, BI explores several technological tools, producing reports and forecasts, in order to improve the efficiency of the decision making process. Such tools include Data Warehouse (DW), Extract-Transform and Load (ETL), Online Analytical Processing (OLAP), Data Mining (DM), Text Mining, Web Mining, Data Visualization, Geographic Information Systems (GIS), and Web Portals.
- On the next level there is a concern with the integration of business processes on BI (Eckerson, 2009; Golfarelli, Rizzi & Cella, 2004; Turban, Sharda, Aroson & King, 2008; Wormus, 2008; Zeller, 2007). According to this approach, “BI is a mechanism to bridge the gap between the business process management to the business strategy” (Zeller, 2008 - pp 3). In addition to all the tools in traditional BI, tools such as Business Performance Management (BPM), Business Activity Monitoring (BAM), Service-Oriented Architecture (SOA), Automatic Decision Systems (ADS), and dashboards are included.
- Adaptive Business Intelligence is concerned with self-learning adaptive systems, that can recommend the best actions, and that could learn with previous decisions, in order to improve continuously (Michalewicz, Schmidt, Michalewicz & Chiriac, 2007). Artificial Intelligence is, in this manner, incorporated into BI systems.

A schematic view of the main approaches that are presented in the literature is depicted in Figure 4. The presented framework can be used as the basis for subsequent research, since it helps to operationalize the current state of the art. Research could be developed along all the presented levels since there are open issues in all of them. Research areas on BI could include integration issues, analysis of usability, assessment, return on investment, and technological issues.

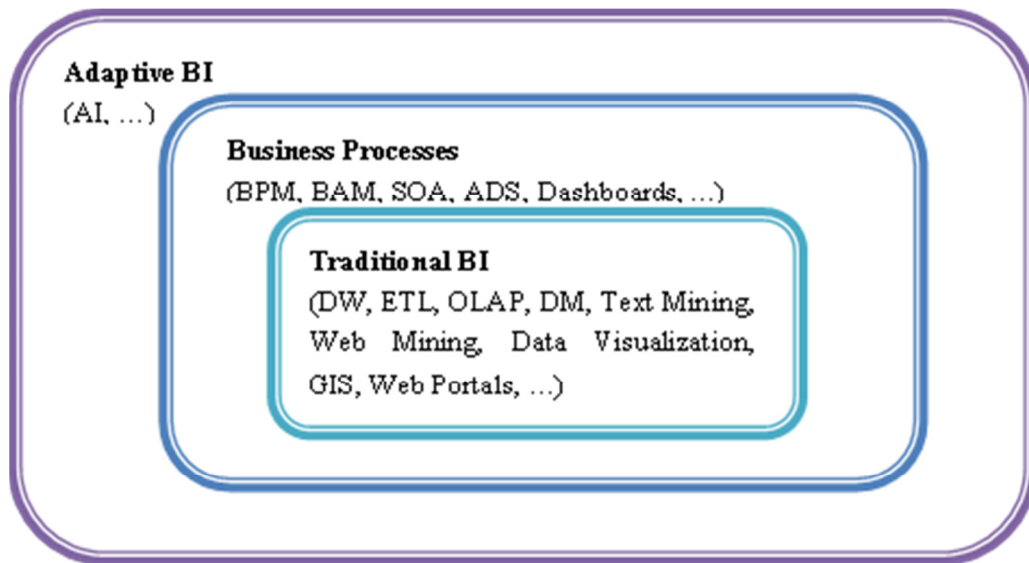


Figure 4 – A Framework for Business Intelligence

2.4 An Example of a Business Intelligence System

The research presented in this thesis focus mainly in traditional BI. A prototype of a BI system was developed in order to support the research. The high-level architecture of the implemented BI system is presented in Figure 5. The underlying relational databases refer to a Higher Education Institution (HEI). Almost all the business processes of the referred HEI are supported by an operational information system, built upon relational databases (Pereira, Azevedo & Castilho, 2007). The direction board of the institution intends to expand the system with the inclusion of a BI system. The architecture includes an ETL module which consolidates data in a data warehouse (DW), a BPM module which helps with the definition of business processes and respective metrics, and an OLAP module which allows for data manipulation.

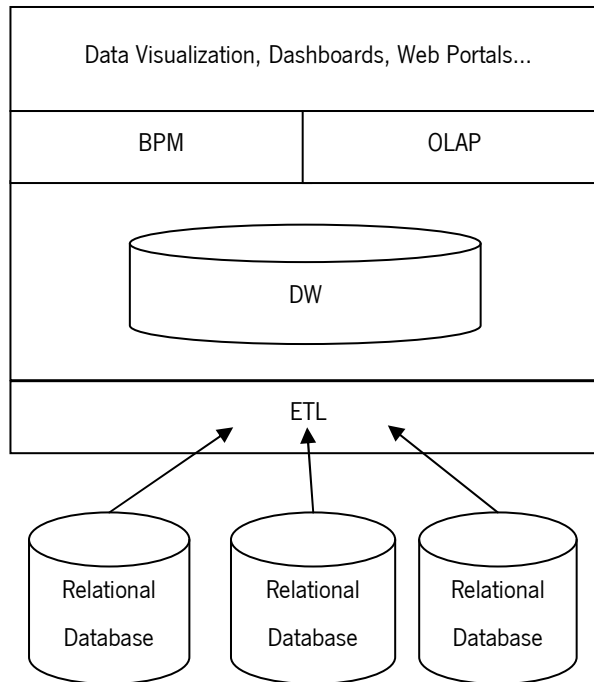


Figure 5 – High-level architecture of a Business Intelligence System

In a first moment, it is intended to deliver information about students, and plans are being done in order to include teachers and employees in consequent phases. Some examples of business questions that can be posed to the system by the HEI responsible, and that can be answered by the BI system, are presented in Table 1. Data visualization tools can help to understand some important issues related with these questions, obtaining answers that support the decision process. OLAP tools are also important as a way to obtain answers to the wide variety of ad-hoc questions posed to the system by business users, considering different dimensions, such as season, time or program. In addition, all these questions can be converted into queries in any of the query languages offered by the Relational Database Management System (RDBMS), for instance SQL or QBE.

Table 1 – Examples of business questions that can be answered by the Business Intelligence System

Question	Dimensions
Who are the best students?	By Season
Who are the worst students?	By Time
How many students conclude the grades according to initial schedule?	By Geography
Which are the courses with higher retention taxes?	By Program
How many students are there?	By Course
...	...

3 Data Mining and Knowledge Discovery in Databases

The term knowledge discovery in databases or KDD, for short, 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, Piatetski-Shapiro & Smyth, 1996). Fayyad considers DM as one of the phases of the KDD process. The DM phase concerns, mainly, the means by which the patterns are extracted and enumerated from data. The literature is sometimes a source of some confusion because the two terms are indistinctly used, making it difficult to determine exactly each of the concepts (Benoît, 2002). Nowadays, the two terms are, usually, indistinctly used and so it will be along this text.

3.1 The Knowledge Discovery in Databases Process

The KDD process, as presented in (Fayyad, Piatetski-Shapiro & Smyth, 1996), is the process of using DM methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database. There are five stages considered, namely, selection, preprocessing, transformation, data mining, and interpretation/evaluation as presented in Figure 6:

- Selection - this stage consists on creating a target data set, or on focusing in a subset of variables or data samples, on which discovery is to be performed;
- Preprocessing - this stage consists on the target data cleaning and preprocessing in order to obtain consistent data;
- Transformation - this stage consists on the transformation of the data using dimensionality reduction or transformation methods;
- Data Mining - this stage consists on the searching for patterns of interest in a particular representational form, depending on the DM objective (usually, prediction);
- Interpretation/Evaluation - this stage consists on the interpretation and evaluation of the mined patterns.

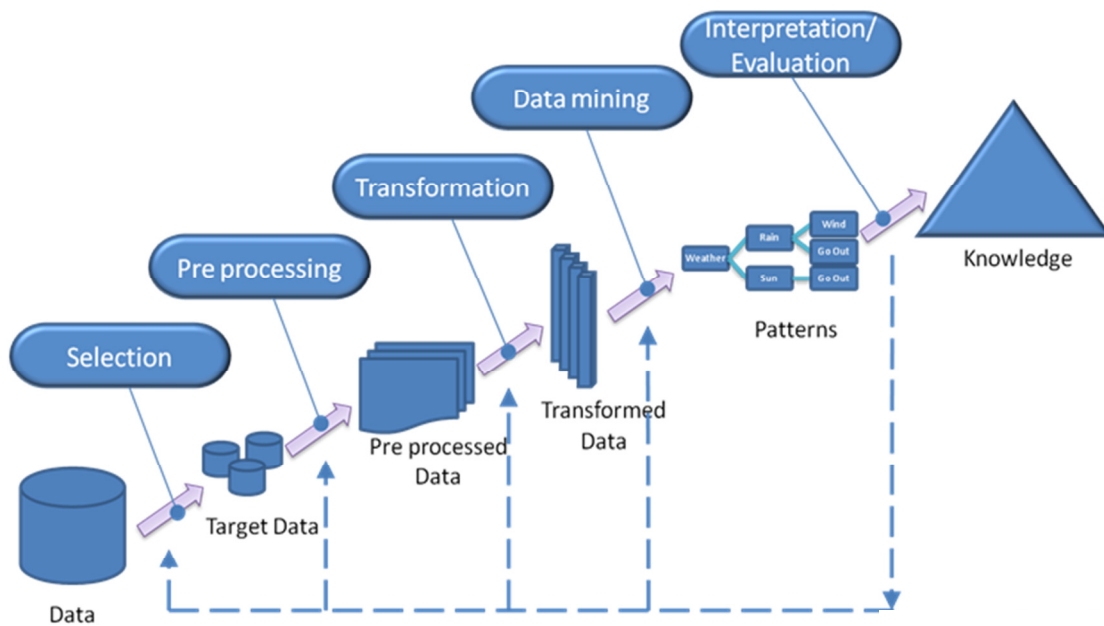


Figure 6 – The KDD Process

The KDD process is preceded by the development of an understanding of the application domain, the relevant prior knowledge, and the goals of the end-user. It must be continued by knowledge consolidation, incorporating this knowledge into the system. The KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user (Brachman & Anand, 1996).

3.2 Some Applications

As of the foundations of KDD and DM, several applications were developed in many diversified fields. The growth of the attention paid to the area emerged from the rising of big databases in an increasing and differentiated number of organizations. Nevertheless, there is the risk of wasting all the value and wealthy of information contained in these databases, unless the adequate techniques are used to extract useful knowledge (Chen, Han & Yu, 1996; Fayyad, 1996; Simoudis, 1996). The application of DM techniques with success can be found in a wide and diversified range of applications, for instance, bioinformatics, ecology and sustainability, finance, industry, marketing, scientific research, telecommunications, and other applications.. Being aware that they are not completely covered here, a great variety of examples are presented in the following lines.

Bioinformatics

In (Salzberg, 1999) DM techniques are applied to gene discovery in DNA sequences. In (Gao, Cao, Qi & Hu, 2005) the authors' work focus also on DNA sequences. In (Gurbaxani & Mallick, 2005), and in (Martin, Gibrat & Rodolphe, 2005) DM is applied to the study of protein structure. In (Wu, Yu & Jang, 2005) the authors developed a framework that helps in the detection of depressive symptoms in psychiatry. In (Chiang, Shieh, Hsu & Wong, 2005) a medical decision support system that can identify the patients who have polyps is presented, while in (Fung & Stoeckel, 2007) the presented work helps on the identification of Alzheimer's disease. An intelligent decision support system to support intensive care medical activities is presented in (Gago & Santos, 2008; Gago, Fernandes, Pinto & Santos, 2009; Santos, Pereira & Silva, 2005).

Ecology and sustainability

In (Silva, Câmara & Escada, 2009) a study of the deforestation problem in Amazonia is presented. In (Nlenanya, 2009) the authors shows the application of a geographical information system (GIS) based on a knowledge discovery interface that can be used to stimulate sustainable development in the sub-Saharan African region. The work presented in (Tadesse, Wardlow & Hayes, 2009) focuses on monitoring and predicting drought's impact on vegetation conditions. A system able to determine if an area is going to be contaminated or not after an oil spill is presented in (Corchado, Mata, Paz & Pozo, 2008). In (Santos, Cortez, Quintela, Neves, Vicente & Arteiro, 2005) DM techniques are used to do the automatic assessment of dam water quality.

Finance

Concerning the applications of DM to finance, in (John, Miller & Kerber, 1996) the authors refer to stock selection in order to obtain the best stock portfolio for investors. Other type of application, concerns the discovery of insurance risks (Apte, Grossman, Pednault, Rosen, Tipu & White, 1999). DM techniques can also be used in credit card fraud detection (Chan, Fan, Prodromidis & Stolfo, 1999). In (Hu, 2005) study, DM is used for analyzing retail banking customer attrition. A study involving bankruptcy prediction based on data mining techniques is presented in (Santos, Cortez, Pereira & Quintela, 2006).

Industry

DM can be applied in the prediction of the ideal moment to replace aircraft components (Létourneau, Famimi & Matwin, 1999). In (König & Gratz, 2005) an application is created in order to optimize manufacturing processes in the semiconductor industry. A real-time decision support system for civil engineering structures is presented in (Quintela, Santos & Cortez, 2007). There are DM applications even in sports: in (Bhandari, Colet, Parker, Pines, Pratap & Ramanujam, 1997) it is described an application used by the National Basketball Association (NBA) coaching staffs to discover interesting patterns in basketball game data.

Marketing

It can be said that some of the most popular applications of DM are in the Marketing field. Applications include market basket analysis and customer segmentation (Ghosh & Strehl, 2005; Simoudis, 1996). In (Hsu, Chung & Huang, 2004) a real data mining application is proposed for personalized shopping recommendation. In (Luck, 2009) DM techniques are applied on CRM data. In (Dzieciolowski & Kina, 2008) it is examined how data mining can help identify the best geographic areas for customer acquisition campaigns. A KDD approach was used to database marketing projects in (Pinto, Gago & Santos, 2006; Pinto, Santos & Marques, 2009; Santos, Cortez, Quintela & Pinto, 2005).

Scientific research

An important application in astronomy is presented in (Fayyad, Djorgovski & Weir, 1996). An approach to geologic study of remotely sensed images is presented in (Smyth, Fayyad, Burl & Perona, 1996). In (Bollacker, Lawrence & Giles, 2000) DM techniques are used to help finding useful publications on the Web. In (Lin, Pu & Lee, 2005) it is presented an application to analyze satellite images. Applications on the social sciences area are presented in (Scime, Murray, Huang & Brownstein-Evans, 2008).

Telecommunications

In (Ezawa & Norton, 1996) DM techniques are applied in order to predict uncollectable telecommunications invoices. In (Pan, Yang, Yang, Li, Li & Li, 2007) the authors aim to identify customers who might switch to a competitor service provider. The problem of telephone calling fraud detection is presented in (Cox, Eick, Wills & Brachman, 1997). The problem of detecting

cellular cloning fraud based on a database of call records is made in (Fawcett & Provost, 1997).

Other applications

New applications appear daily. In (Lappas, 2009) several applications are presented to societal benefit areas such as helpdesks and recommendation systems, digital libraries, e-learning, security and crime investigation, e-government services, e-politics and e-democracy. In (Rahman, 2009) a few areas to data mining applications are highlighted.

3.3 Data Mining Tasks, Methods/Algorithms, and Models/Patterns

Prediction and description were identified as the two “high-level” primary goals of DM (Fayyad, Piatetski-Shapiro & Smyth, 1996). “Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest. Description focuses on finding human-interpretable patterns on finding the data.” (Fayyad, Piatetski-Shapiro & Smyth, 1996 - pp 12)

To achieve these goals some tasks were used and its description can be found in the literature. Some of the most common tasks are classification, prediction, clustering, association, and summarization:

- Classification consists in finding a function that associates an instance of the independent variables to a specific pre-defined value of the target variable, named as class. The target variable should be of nominal type;
- Prediction consists in finding a function that associates an instance of the independent variables to some numerical value of a real-valued target variable, in order to predict future unknown values for that target variable;
- Clustering allows the identification of homogeneous groups containing several elements which have high similarity with all the other elements of the same group, and that have low similarity to all the elements of the other groups;
- Association consists in finding a model that describes significant dependencies between variables, that is to say, identifying facts that can be directly or indirectly associated;

- Summarization uses methods to discover a compact description for data, in order to find a better description of the data, thus improving its understanding.

A significant number of methods/algorithms have been developed to accomplish each task, and different kinds of models/patterns can be obtained. Classification methods include decision trees, classification rules, neural networks, support vector machines, Bayesian data analysis, Bayesian networks, and k-nearest neighbor. Prediction methods include linear regression, nonlinear regression, neural networks, decision trees, and k-nearest neighbor. Clustering methods include partitioning, hierarchical and model-based methods. Association is accomplished with association rules. Summarization methods include EDA² and OLAP³.

- Decision trees can be defined as a way to represent a group of rules that follow a hierarchy. Each node of the tree defines a test for some variable, and each leaf defines a class. Three of the most popular algorithms to generate decision trees are ID3⁴, C4.5, and CART⁵ that use a divide-and-conquer approach to generate the rules. ID3 and C4.5, which can be considered an improvement of ID3, use the concept of information gain and CART uses the concept of Gini index as guidance;
- Classification rules are IF-THEN expressions of the form *IF antecedent THEN consequent*. The antecedent is formed by a group of tests for the target variables, and the consequent defines the class that satisfies those conditions. Rules can be obtained directly from decision trees. An alternative approach is using a different way to deal with the situation, known as covering or separate-and-cover algorithms. Two examples of such algorithms are PRISM and RIPPER⁶, using coverage and accuracy as guidance;
- Neural networks intend to simulate the human brain. A neural network consists in a computational structure based in processing units, the neurons, communicating by sending signals through links. Each neuron possesses inputs and outputs each associated with a weight. To build the neural network the number of neurons as well as the weights of each link must be discovered through the training of the network. Algorithms include multilayer perceptron and back propagation. One of the main

² Exploratory Data Analysis

³ On-Line Analytical Processing

⁴ Iterative Dichotomiser 3

⁵ Classification and Regression Trees

⁶ Repeated Incremental Pruning to Produce Error Reduction

disadvantages appointed to neural networks are based on that they are similar to black boxes since their internal structure is unknown;

- Support vector machines are originated from research in the area of statistical learning. Support vector algorithms “select a small number of critical boundary instances called support vectors from each class and build a linear discriminant function that separates them from each class and build a linear discriminant function that separates them as widely as possible.” (Witten & Frank, 2005 – pp 188). Support vector machines are closely related to neural networks;
- Bayesian data analysis is based on Bayes theorem of conditional probability. It consists in obtaining a probability distribution from the observed data, starting with a joint probability distribution, and then computing the posterior probability from this prior probability, using Bayes theorem. The main disadvantage is that it involves heavy calculations;
- Bayesian networks are used to represent knowledge from an uncertain domain. They are direct acyclic graphs whose nodes represent random variables of interest, and whose edges represent the conditional dependencies between the variables. K2 and TAN⁷ are two specific algorithms to learning bayesian networks;
- K-nearest neighbor is a simple and popular classification method. With nearest-neighbor “each new instance is compared with existing ones using a distance metric and the closest existing instance is used to assign the class to the new one. (...) Sometimes more than one nearest neighbor is used, and the majority class of the closest k neighbors (or the distance-weighted average, if the class is numeric) is assigned to the new instance. This is termed k-nearest-neighbor method.” (Witten & Frank, 2005 – pp 78);
- Linear regression intends to discover a function that represents an approximate behavior of numerical variables, by expressing the target or dependent variable as a linear combination of the other variables, named as the independent variables. During the training the weights are calculated so that the differences between the real values and the predicted ones are minimized;

⁷ Tree-Augmented Naive Bayes

- Nonlinear regression is similar to linear regression, but instead of using a linear combination of the independent variables to predict the value for the dependent variable, it uses a nonlinear function of the independent variables to obtain a prediction of the dependent variable;
- Partitioning methods for clustering, results in a fixed number of mutually exclusive groups, named as clusters. Each cluster is represented by one of his members, named as centroid. One of the main issues related with this kind of method is the determination of the ideal number of cluster. Traditional algorithms include k-means and k-medoids;
- Hierarchical methods, instead of returning an unstructured set of clusters, return a hierarchical tree structure, a dendrogram, which defines a hierarchy of clusters. A measure of dissimilarity is required and it is not necessary to predefine the number of clusters. Algorithms include BIRCH⁸, ROCK⁹, and Chameleon;
- Model-based clustering also named as probability-based clustering, are the clustering methods most closely to statistics. It is assumed that clusters are represented by a mixture of probability distributions and some methods are used to find the parameters. This methods aims at overcoming some of the issues related to other clustering methods. The EM¹⁰ algorithm is one of the most known algorithms using this technique;
- Association rules are IF-THEN expressions of the form *IF antecedent THEN consequent*. They are created by analyzing data for frequent IF/THEN patterns and then identifying the most relevant, interesting, and useful ones. To select those most relevant, interesting and useful rules from the set of all possible rules, various measures of significance and interestingness can be used, usually support and confidence. Apriori, GRI¹¹, and FP-growth¹² are examples of classical algorithms used to generate association rules.
- EDA, “as the name suggests, the goal here is simply to explore the data without any clear ideas of what we are looking for. Typically, EDA techniques are interactive and

⁸ Balanced Iterative Reducing and Clustering

⁹ RObust Clustering using linKs

¹⁰ Expectation-Maximization

¹¹ Generalized Rule Induction

¹² Frequent Pattern Growth

visual, and there are many effective graphical display methods for relatively small, low-dimensional data sets.” (Hand, Mannila & Smyth, 2001 – pp 11)

- OLAP tools provide environments for advanced data analysis, doing the synthesis, analysis, and consolidation of big volumes of data, stored in a multi-dimensional perspective. This multidimensional perspective allows the analysis of business data according to several dimensions. It is very important that OLAP tools provide user friendly interfaces, in order to become much more useful to data analysis.

An outline of this DM tasks, methods/algorithms, models/patterns, and guidance is presented in Table 2.

There are different forms of evaluating models’ interestingness in each case, such as cross-validation, bootstrapping, bagging and boosting, estimating confidence intervals, or ROC curves. There are also a large variety of alternatives to provide guidance, including accuracy and error measures. We will not discuss in more detail each one of these issues, since we consider it is outside the scope of this text. Several textbooks can be found that cover these topics in more detail, e.g. (Han & Kamber, 2006; Hand, Mannila & Smyth, 2001; Larose, 2005; Myatt, 2007; Santos & Azevedo, 2005; Witten & Frank, 2005; Ye, 2003).

The emergence of more complex types of data led to the development of new methods and models to cope with the new task of mining complex data. As examples, we can point out text mining (Prado & Ferneda, 2008), web mining (content, structure, and usage) (Markov & Larose, 2007), spatial data mining (Nlenanya, 2009), graph mining (Zhang, Hu, Xia, Zhou & Achananuparp, 2008), mining time-series data (Liabotis, Theodoulidis & Saraaee, 2006), among others. In (Kumar, 2011) some trends and new domains are explored.

Data mining languages for business intelligence

Table 2 – Outline of DM tasks, Methods/Algorithms, Models/Patterns, and Guidance

DM Tasks	Methods/Algorithms	Models/Patterns	Guidance
Classification	Decision trees (ID3, C4.5, CART)	Tree	Information Gain / Gini Index
	Classification rules (PRISM, RIPPER)	Rules	Accuracy and Coverage
	Neural Networks (multilayer perceptron, back propagation)	Neural Network	Error
	Support Vector Machines	Maximum margin hyperplane	Error
	Bayesian data analysis	Probability distribution	Conditional Probability
	Bayesian networks (K2, TAN)	Directed acyclic graph	Conditional Probability
	k-nearest neighbor	Pattern space	Distance function
Prediction	Linear regression	Linear function	Error
	Nonlinear regression	Nonlinear function	Error
	Neural Networks (multilayer perceptron, back propagation)	Neural Network	Error
	Decision trees (ID3, C4.5, CART)	Tree	Information Gain Gini Index
	k-nearest neighbor	Pattern space	Distance function
Clustering	Partitioning (k-means, k-medoids)	Diagram	Measure of dissimilarity
	Hierarchical (BIRCH, ROCK, Chameleon)	Diagram	Measure of dissimilarity
	Model-based (EM, Kohonen networks)	Diagram	Measure of dissimilarity
Association	Association rules (Apriori, GRI, FP-growth)	Rules	Support and Confidence
Summarization	EDA	Tables, Charts	—
	OLAP	OLAP cubes	—

4 Data Mining Languages

Some efforts are being made seeking the establishment of standards in the DM area, both by academics, and by people in the industry field. The main goal is to integrate DM with relational databases, thus allowing an easier application of DM to business systems, and making it more available to decision making. An important issue in this domain concerns data mining languages.

4.1 Towards Standards for Data Mining

Examining the primary conferences and journals in the DM field, it can be concluded that the main issues for research are related to improving data preparation for data mining, to developing better algorithms and methods for specific problems and applications, and to measuring the utility of the obtained models. Nevertheless, the necessity to develop a theory for DM, similar to the one that was developed by Codd, with the Relational Model for database systems, arose (De Raedt, 2003; Imielinski & Mannila, 1996; Mannila, 2000). Over the past few years, some efforts have been made in the development of standards for DM and KDD (Dzeroski, 2007; Mannila, 2000). These efforts arise both from academics and from people in the industry field. Being aware that they may be not completely covered here, the authors present the ones that they consider to be most important. Industrial standards are presented in section 4.1.1, and scientific research is presented in section 4.1.2.

4.1.1 Industrial Standards

Some of the efforts in the industrial field concern the definition of processes/methodologies that can guide the implementation of DM applications. For instance, SEMMA and CRISP-DM can be pointed out as such examples. In (Azevedo & Santos, 2008) a comparative study of these processes is presented.

The acronym SEMMA stands for Sample, Explore, Modify, Model, Assess, and refers to the process of conducting a DM project. The SAS Institute considers a cycle with 5 stages for the process, which are, sample, explore, modify, model, and assess:

- Sample - this stage consists on sampling the data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly;

- Explore - this stage consists on exploring the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas;
- Modify - this stage consists on modifying the data by creating, selecting, and transforming the variables to focus the model selection process;
- Model - this stage consists on modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome;
- Assess - this stage consists on assessing the data by evaluating the usefulness and reliability of the findings from the DM process and estimate how well it performs.

SEMMA offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects. It thus confers a structure for his conception, creation and evolution, helping to present solutions to business problems as well as to find the DM business goals (Santos & Azevedo, 2005).

CRISP-DM stands for Cross-Industry Standard Process for Data Mining. It consists on a cycle that comprises six phases, which are business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Figure 7):

- Business understanding - this initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives;
- Data understanding - the data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information;
- Data preparation - the data preparation phase covers all activities to construct the final dataset from the initial raw data;
- Modeling - in this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values;
- Evaluation - at this stage the model (or models) obtained are more thoroughly evaluated and the steps executed to construct the model are reviewed to assure it properly achieves the business objectives;

- Deployment – the creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.

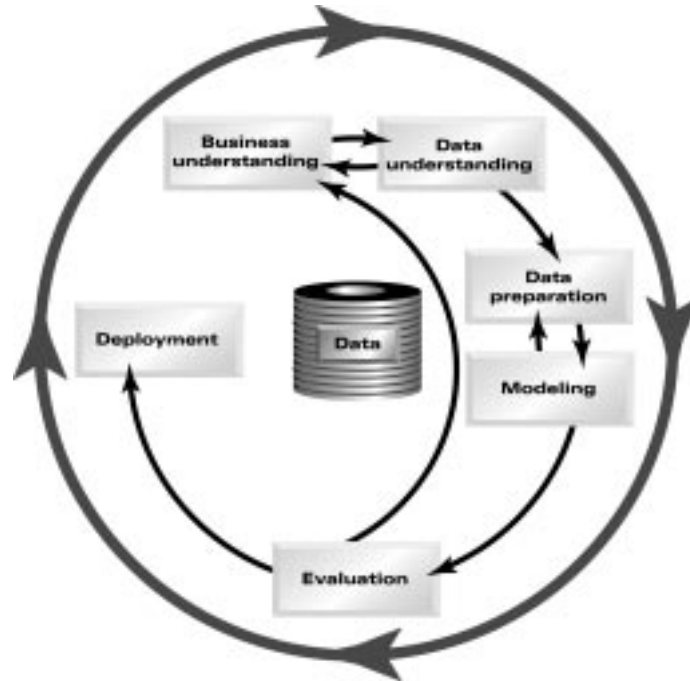


Figure 7 - The CRISP-DM life cycle (Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer & Wirth, 2000 - pp 13)

CRISP-DM is extremely complete and documented. All its stages are duly organized, structured and defined, allowing a project to be easily understood or revised (Santos & Azevedo, 2005).

By doing a comparison between the KDD and SEMMA stages we would, on a first approach, state that they are equivalent: *Sample* can be identified as *Selection*, *Explore* can be identified as *Pre-processing*, *Modify* can be identified as *Transformation*, *Model* can be identified as *DM*, *Assess* can be identified as *Interpretation/Evaluation*. Examining it thoroughly, we can state that the five stages of the SEMMA process can be seen as a practical implementation of the KDD process five stages, since it is directly linked to the SAS Enterprise Miner software.

Comparing the KDD stages with the CRISP-DM stages is not as straightforward as in the SEMMA situation. Nevertheless, we can first of all observe that the CRISP-DM methodology incorporates the steps that, as mentioned above, must precede and follow the KDD process; that is to say, an understanding of the application domain, the relevant prior knowledge, and the goals of the end-user, which must precede KDD, and knowledge consolidation, which must follow KDD. The *Business Understanding* phase can be identified as the development of an

understanding of the application domain, the relevant prior knowledge and the goals of the end-user. The *Deployment* phase can be identified as the knowledge consolidation. Concerning the remaining stages, we can say that: The *Data Understanding* phase can be identified as the combination of *Selection* and *Pre-processing*; the *Data Preparation* phase can be identified as *Transformation*; the *Modeling* phase can be identified as *DM*; the *Evaluation* phase can be identified as *Interpretation/Evaluation*. In Table 3 a summary of these correspondences is presented.

Considering the presented analysis we conclude that SEMMA and CRISP-DM can be seen as implementations of the KDD process described in (Fayyad, Piatetski-Shapiro & Smyth, 1996). At first sight, we can get to the conclusion that CRISP-DM is more complete than SEMMA. However, in a deeper analysis, we can integrate the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user, on the Sample stage of SEMMA; this can be done because the data cannot be sampled unless there is a truthful understanding of all the presented aspects. With regarding consolidation, by incorporating this knowledge into the system, we can assume that it is included, because it is truly the reason for doing it. This leads to the fact that standards have been achieved, concerning the overall process: SEMMA and CRISP-DM do guide people to know how DM can be applied in practice in real systems.

Table 3 – Summary of the correspondences between KDD, SEMMA and CRISP-DM

KDD	SEMMA	CRISP-DM
Pre KDD	———	Business understanding
Selection	Sample	Data Understanding
Pre processing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	———	Deployment

Other efforts in the industrial field focus on the development of software suites for implementing some selected DM algorithms. Over the past few years several data mining suites have been developed (KDNuggets, 2011). These suites deliver user-friendly environments that allow users to apply data mining freely and easily. Some of them have capabilities to integrate all the KDD process. Nevertheless, if these suites are used without the knowledge of a DM expert, the obtained results may not be useful. This is due to the fact that all the KDD process must be considered, in spite of just applying DM algorithms without being aware of their characteristics. In addition, these suites are oriented to DM specialists and not to business users.

There are some efforts being made that intend to develop standards that will allow model representation to be platform independent. Such an example is the Knowledge Discovery Metamodel (KDM) (Object Management Group, 2008). Another example is the Predictive Model Markup Language (PMML) (Data Mining Group, 2009). OLE DB for Data Mining can also be presented as an example (Tang & MacLennan, 2005). These models mainly seek portability among models obtained in different tools, and some of them are included in most of the BI tools in the market.

4.1.2 Scientific Research

Above all the academic efforts towards a theory for DM and KDD follow closely the theory developed by Codd for the Relational Model. According to Codd's Relational Model (Codd, 1970; Codd, 1982), a database consists on a set of *relations*. Each *relation* is a set of *tuples*. Two formal languages were defined: the *Relational Algebra* and the *Relational Calculus*. A fundamental property of such languages is *closure*. A very popular language implemented nowadays in all the RDBMS is SQL. Research towards a theory for DM and KDD focuses mainly on obtaining a theory, similar to Codd's theory, giving DM and KDD a database perspective. In (Imielinski & Mannila, 1996) the actual *DM systems* are compared to *File Systems*. Thus, the authors refer the aim of developing Knowledge and Data Discovery Management Systems (KDDMS), as being similar to the existing RDBMS.

A promising research line is that of *Inductive Databases*, as presented by (De Raedt, 2003) and (Imielinski & Mannila, 1996). According to the Inductive Databases framework, data and models are stored on the same database and can be queried at the same level. Based on this framework, some theoretical research and prototypes have been developed, as well as some research about Data Mining Languages. A different perspective is given in (Catania,

Maddalena, Mazza, Bertino & Rizzi, 2004), who present a framework that keeps data and patterns separated. An algebra for DM, the K-algebra, is presented by (Gerber & Fernandes, 2004). A more theoretical approach is the 3W-model presented in (Calders, Lakshmanan, Ng & Paredaens, 2006), which is an extension of the relational algebra. (Nijssen & De Raedt, 2007) presents IQL, an extension of the relational calculus.

4.2 Inductive Databases and Data Mining Languages

“Inductive databases tightly integrate databases with data mining. The key ideas are that data and patterns (or models) are handled in the same way and that an inductive query language allows the user to query and manipulate the patterns (or models) of interest” (De Raedt, 2003, pp 69).

Inductive databases research goal is to achieve the entire KDD process by the use of queries to an inductive database. Besides the traditional queries of the relational model, there is the necessity to consider inductive queries that will be used to generate and manipulate the DM models (Dzeroski, 2007). This can be provided in several distinct ways, and thus many different research lines can be found. For instance, in (Mielikäinen, 2004) the authors try to clarify what distinguishes traditional databases from inductive databases, arguing that it is the ability of the second to rank or to grade queries. According to this line of research, queries consist of constraints and the aim is to develop a language of patterns and a set of constraints that patterns must satisfy.

An inductive database should provide the following features (Bonchi, Giannotti, Lucchesse, Orlando, Perego & Trasarti, 2007):

- Coupling with a database management system:
 - Capability for retrieving data of interest;
 - Data and patterns on the same DBMS;
- Expressiveness of the query language:
 - High-level vision of the pattern discovery system similar to the high-level vision of the DBMS;
- Efficiency of the mining engine:
 - Capability for efficient query response time;
- Graphical user interface:
 - Capability for pattern visualization and existence of navigation tools.

The Inductive query language is a fundamental issue to consider in the research. Two different approaches can be found in the literature:

- Definition of special purpose languages, presented in section 4.2.1;
- Using just standard SQL, presented in section 4.2.2.

4.2.1 Special Purpose Languages

The definition of special purpose languages was the research line chosen by several researchers. There is a line of research focusing on the definition of SQL extensions. There are also some languages based on XML. Logic-based languages can also be found. Several approaches are included in the literature. Following, some examples considered as the most relevant are presented, SQL-based languages in section 4.2.1.1, XML-Based languages in section 4.2.1.2, and Logic-Based languages in section 4.2.1.3.

4.2.1.1 SQL-Based Languages

A Data Mining Query Language, DMQL, is presented in (Han, Fu, Wang, Koperski & Zaiane, 1996). DMQL has a similar syntax to that of SQL. It allows defining the data to be mined, the kind of knowledge to be discovered, the inclusion of background knowledge, and the definition of thresholds. The kind of knowledge to be mined concerns different types of rules, for instance, association rules and classification rules.

Another approach is presented in (Meo, Psaila & Ceri, 1998). The *MINERULE* operator, which is an extension of SQL and has got a similar syntax to that of SQL, is presented. The operator mines for association rules, allowing the definition of groups to which mining is applied.

MSQL is presented in (Imielinski & Virmani, 1999). The language also has got similar syntax to that of SQL, and mines for rules. MSQL has got two main commands, namely, *GetRules* and *SelectRules*. *GetRules* generates rules from data and *SelectRules* queries a pre-existing database. The problem of providing little support to the pre-processing and pos-processing phases of the KDD process is common to all of these languages (Botta, Boulicaut, Masson & Meo, 2004).

A language supporting pre- and post- processing phases is presented in (Kramer, Aufschild, Hapfelmeier, Jarasch, Kessler, Reckow, Wicker & Richter, 2006), and it is a component of SINBAD system. It consists of an extension of SQL, and several operators are defined. For instance, the operator *extend add as* is used to add the results of data mining operations as new attributes to a relation, and the operator *feature select* allows the selection of tuples in a

relation by defining specific conditions. The authors sustain that the defined language can handle the pre-processing techniques discretization and feature selection, as well as the data mining techniques pattern discovery, clustering and classification, but there is no clear indication about the supported models.

SPQL (Simple Pattern Query Language) is presented in (Bonchi, Giannotti, Lucchesse, Orlando, Perego & Trasarti, 2007). The language has got similar syntax to that of SQL, while the mining is made with the clause TRANSACTION. The language mines for frequent patterns, and handles the pre-processing phase. The language serves as the base for a complete constraint based querying system, ConQuesSt, which is a human-guided, interactive and iterative system for pattern discovery.

Just to give a glance at the syntax of some of the presented data mining query languages, an example is given in Table 4 that allows a comparison between them. The language that is a component of the SINBAD system is not included because there is no sufficient information about the language syntax. SPQL is not included since this language does not allow classification rules.

Analyzing the presented languages, it can be concluded that all of them have limitations on the types of models they support, and that more research is needed in this area.

Table 4 – Comparison of SQL-Based languages syntax

<p>Schema: student(id,gender,age,nenroll,grant,grade)</p> <p>Classification Rules for grade in consequent</p> <p>Having grade<10; support>0.1; confidence>0.2</p>	
DMQL	<p>use database school</p> <p>find classification rules as Classification Rules</p> <p>according to grade</p> <p>Related to gender, age, nenroll, grant</p> <p>From student</p> <p>Where student.grade<10</p> <p>With support threshold > 0.1</p> <p>With confidence threshold > 0.2</p>
MineRule	<p>MINE RULE ClassificationRules AS</p> <p>SELECT DISTINCT gender, age, nenroll, grant AS BODY, grade AS HEAD</p> <p>FROM student</p> <p>WHERE grade<10</p> <p>EXTRACTING RULES WITH SUPPORT: 0.1, CONFIDENCE: 0.2</p>
MSQL	<p>GetRules (student)</p> <p>Into ClassificationRules</p> <p>Where consequent is {(grade<10)}</p> <p>and body in {(gender=*), (age=*), (nenroll=*), (grant=*)}</p> <p>and confidence > 0.2</p> <p>and support > 0.1</p>

4.2.1.2 XML-Based Languages

KDDML, which stands for KDD Markup Language, consists of a middleware language and system, as expressed by the authors in (Romei, Ruggieri & Turini, 2006). The language is entirely based in XML standards, including query syntax, data, and model representations. Queries consist of XML documents and operations consist of XML tags. According to the presented examples, the kinds of models that are dealt by the system are trees, clusters and rules.

Another example of an XML-based system, named XDM (XML for Data Mining), is presented in (Meo & Psaila, 2006). The basic idea consists of definitions of two concepts: *Data Item*, which

is a data/patterns container, and *Statement*, which is a description of an operator application. The aim of the system is the adoption of XML in the inductive database framework. The presented examples include association rules and clusters.

4.2.1.3 Logic-Based Languages

In (De Raedt, 2002) it is presented a constraint logic programming language, named RDM, which stands for Relational Database Mining, developed to support DM. The language is embedded within Prolog. In this research, examples are presented for association rules and experiments were made on graph structures.

4.2.2 Languages Using Standard SQL

Several researchers chose a different approach. Using this approach, the inductive database can be queried using standard SQL. This approach has got advantages over the approach of using special-purpose DM languages concerning extensibility and flexibility. An example, is the research presented in (Boulicaut, Klemettinen & Mannila, 1999), where the principle is demonstrated for association rules. In (Sarawagi, Thomas & Agrawal, 2000) the same principle is used considering that association rules and performances of several alternatives are compared, by means of distinct SQL versions (SQL92 and SQL-OR). Using only basic SQL3 constructions and functions, Jamil shows that any object relational database can be mined for association rules (Jamil, 2004). In (Rantzau, 2004) approaches based on SQL-92 are investigated, and a new approach named Quiver is presented; this approach employs universal and existential quantifiers to find frequent itemsets. In (Calders, Goethals & Prado, 2006) the authors propose extensions of RDBMS and introduce the notion of *virtual mining views*, which can be queried since they are traditional relational relations (views). Using association rules and frequent itemsets as an example, they show that the user can query mining results by using only SQL. Trying to overcome the burden of the use of a limited type of models, in (Fromont, Blockeel & Struyf, 2007) the authors investigate how this approach can be used for models such as decision trees.

4.3 Data Mining Integration with Relational Databases

The presented languages are part of bigger projects that intend to develop a complete system in order to incorporate the entire KDD process. The same goes for the projects using standard SQL. The common aim of all the presented projects and, in general, of research in the area of

inductive databases is, undoubtedly, to achieve the KDDMS referred in (Imielinski & Mannila, 1996), that allows the high-level abstraction present on the RDBMS, and that integrates the complete KDD process.

The importance of KDDMS is similar to the importance of RDBMS. RDBMS released users from the burden of becoming aware of the technical details of file systems. This was achieved by means of physical and logical independence between data and applications. This fact allowed final business users to put ad-hoc questions directly to the systems, thus making systems truly available for them.

5 Query-By-Example Languages

Codd's relational model for databases (Codd, 1970) has been adopted long ago in organizations. In relational databases, data is stored in tables also named as relations. A set of relations forms a database. The description of the database is known as the Database Schema. In Figure 8 an example of a database schema is presented. The considered database stores data from students' enrollment in exams and their respective grades in a higher education institution. The success of Codd's relational model for databases led to the development of several languages that allow data manipulation and that also allow obtaining quick answers to ad-hoc business questions through queries on the data stored in databases.

Initially, two formal languages were defined for relational databases: Relational Algebra and Relational Calculus (Codd, 1970; Codd, 1971). Since that time, several languages were developed in order that users could access data stored in databases. Query-By-Example (QBE) languages were developed with success. Since the first developments (Zloof & de Jong, 1977; Zloof, 1977; Zloof, 1975), many advances occurred in the area, and the philosophy behind QBE is being applied in several distinct areas (Braga, Campi, Ceri & Spoletini, 2007; Ferreira, Cruz & Henriques, 2009; Gokhale & Aslandogan, 2003; Malerba, Appice & Vacca, 2002; Papadias & Sellis, 1995; Sweets, Pathak & Weng, 1998). QBE languages are nowadays available in several RDBMS. Those languages allow business users to directly manipulate data without the need of developing programming skills. It can be said that a QBE language is business oriented, and is iterative and interactive in nature since it allows obtaining answers to ad-hoc business questions that can be directly converted into QBE queries. Business users frequently pose business questions that can be answered through queries to a database. Those queries allow the selection of the database's data that provide the answers to the referred business questions. The use of QBE languages by business users in order to directly obtain those answers is an usual practice in organizations nowadays.

STUDENT (<u>StudentID</u> , Student Name, Student Gender, ProgramID, IDCard, IDCard Date, Birthdate, ...)
PLAN (<u>Plan</u> , ProgramID, CourseID, Program Year, Program Semester, #Theoretical, #Practical, ...)
ENROLLMENT (<u>Ref</u> , StudentID, ProgramID, CourseID, Enrollement Date, Year, Season, Semester, ...)
GRADE (<u>Ref</u> , Grade, Version, TeacherID, Date, Validation Date, RegistryID, ...)
REGISTRY (<u>RegistryID</u> , CourseID, ProgramID, Exam Type, Exam Date, Print Date, ...)
TEACHER (<u>TeacherID</u> , Teacher Name, Department, Rank, Qualifications, Birth Date, Admission Date, ...)
COURSE (<u>CourseID</u> , Course Designation)
PROGRAM (<u>ProgramID</u> , Program Designation)

Figure 8 – Relational Database Schema

5.1 General Notions

Query-By-Example are declarative, also called nonprocedural or very high level languages. By using this type of languages the user defines “what s/he wants to do” instead of defining “how to do it”, which is typical of imperative languages. According to Zloof, Query-by-Example is: “a high-level database management language that provides a convenient and unified style to query, update, define, and control a relational database. The philosophy of Query-by-Example is to require the user to know very little in order to get started and to minimize the number of concepts that s/he subsequently has to learn in order to understand and use the whole language.” (Zloof, 1977 – pp 324). In this type of languages, queries are presented in the form of skeleton tables showing, as example, the necessary tables and corresponding columns that are necessary to answer the business questions linked to each query (Figure 9).

Table →				
Column →				
Criteria →				

Figure 9 – Skeleton table for a QBE language

In the first line of the skeleton table, the user indicates which tables contain the necessary data. In the second line of the skeleton table the user indicates which columns, from each of the tables indicated in the first line, contain the necessary data. In the third column of the skeleton table the user will be able to define different criteria corresponding to constraints on the data.

QBE languages allow obtaining answers to ad-hoc business questions. Those business questions are converted into queries to the system, written in QBE language. Following, some basic examples of business questions and corresponding queries in QBE are presented, in order to a better understanding. The examples are based on the relational schema presented in Figure 8. The presented business questions were considered having in mind that different types of queries were involved. Query 1 involves data from only one table in the DB. Query 2 involves data from more than one table in the DB. Query 3 involves only one criterion. Query 4 involves more than one criterion.

Business Question 1

Obtaining the list of teachers' qualifications.

QBE query 1

There is only one table containing the necessary data, namely TEACHER. The necessary columns are: *Teacher name* and *Qualifications*. The query is presented in Figure 10 and the obtained result/answer is presented in Figure 11.

Q1:

Table →	TEACHER	TEACHER		
Column →	Teacher Name	Qualifications		
Criteria →				

Figure 10 – QBE query 1

Obtained Result/Answer 1

Teacher Name	Qualifications		
Margarida Maria Mato	Mestre		
Maria João Maia Pinto	Doutora		
Maria do Carmo Azere	Mestre		
Maria Helena Antunes	Doutora		
Maria Helena da Costa	Mestre		
Maria Helena Salazar c	Mestre		
Mariana Curado Malta	Mestre		
Mário Nuno Ferreira M	Dr.		

Figure 11 – QBE query 1 result/answer

Business Question 2

Obtaining the list of students' grades for each course.

QBE query 2

Tables and corresponding columns containing the necessary data are: column *Course Designation* from table *COURSE*, column *Student Name* from table *STUDENT*, and column *Grade* from table *GRADE*. There are no criteria to be defined. The query is presented in Figure 12 and the obtained result/answer is presented in Figure 13.

Q2:

Table →	COURSE	STUDENT	GRADE	
Column →	Course Designation	Student Name	Grade	
Criteria →				

Figure 12 – QBE query 2

Obtained Result/Answer 2

Course Designation	Student Name	Grade
COMUNICAÇÃO ORGANIZACIONAL I	JOAQUIM PAULO DA CRUZ OLIV	16
LÍNGUA E CULTURA ESTRANGEIRA C I - ESPANHOL	MARTA SUZETE CORREIA ANDR	16
INTERPRETAÇÃO CONSECUTIVA E SIMULTÂNEA I	LILIANA COSTA DA SILVA QUIN	15
PSICOSSOCIOLOGIA DAS ORGANIZAÇÕES	JOAO FRANCISCO DE MAGALHA	15
ESTUDOS INTERCULTURAI	MARIA ANDREIA DA SILVEIRA F	15
DIREITO DAS SOCIEDADES	FILIPPE MIGUEL CRUZ FERNANDE	15
DIREITO DAS SOCIEDADES	ANA RAQUEL ALMEIDA TEIXEIR	15
ESTUDOS INTERCULTURAI	NELSON RICARDO LOURENÇO G	15
PSICOSSOCIOLOGIA DAS ORGANIZAÇÕES	MARIA ISABEL OLIVEIRA MACIE	15
PSICOSSOCIOLOGIA DAS ORGANIZAÇÕES	GAELE CRINE	15

Figure 13 – QBE query 2 result/answer

Business Question 3

Obtaining the list of students’ grades by course, from Season NM and from semester 2.

QBE query 3

Tables and corresponding columns containing the necessary data are: column *Course Designation* from table *COURSE*, column *Student Name* from table *STUDENT*, column *Grade* from table *GRADE*, and columns *Season* and *Semester* from table *ENROLLEMMENT*. There are two criteria to consider: Season = “NM” and Semester = 2. The query is presented in Figure 14 and the obtained result/answer is presented in Figure 15.

Q3:

Table →	COURSE	STUDENT	GRADE	ENROLLEMENT	ENROLLEMENT
Column →	Course Designation	Student Name	Grade	Season	Semester
Criteria →				= NM	= 2

Figure 14 – QBE query 3

Obtained Result/Answer 3

DESIG	NOME	NOTA	TIPO	SEMESTRE
COMÉRCIO ELECTRÓNICO/DIREITO INTERNACIONAL	DANIELA PAIS DE	14	NM	2
SEMINÁRIO/SEMINÁRIO SOBRE EMPREENDEDORISM	DANIELA PAIS DE	24	NM	2
NOÇÕES DE CONTABILIDADE	DANIELA PAIS DE	14	NM	2
OPÇÃO I CONDIÇÃO A	DANIELA PAIS DE	24	NM	2
CONSOLIDAÇÃO DE DEMONSTRAÇÕES FINANCEIRAS	DANIELA PAIS DE	14	NM	2
CONTABILIDADE ANALÍTICA	DANIELA PAIS DE	14	NM	2
CONTABILIDADE ANALÍTICA EXPLORAÇÃO I	DANIELA PAIS DE	14	NM	2
CONTABILIDADE ANALÍTICA EXPLORAÇÃO III	DANIELA PAIS DE	14	NM	2
CONTABILIDADE E ANÁLISE FINANCEIRA	DANIELA PAIS DE	14	NM	2

Figure 15 – QBE query 3 result/answer

Business Question 4

Obtaining the list of students’ grades by course, from Seasons NM and RE, both from semester 1.

QBE query 4

Tables and corresponding columns containing the necessary data are: column *Course Designation* from table *COURSE*, column *Student Name* from table *STUDENT*, column *Grade* from table *GRADE*, and columns *Season* and *Semester* from table *ENROLLEMMENT*. The following criteria must be defined: (Season = “NM” and Semester = 1) or (Season = “RE” and Semester = 1). The query is presented in Figure 16 and the obtained result/answer is presented in Figure 17.

Q4:

Table →	COURSE	STUDENT	GRADE	ENROLLEMENT	ENROLLEMENT
Column →	Course Designation	Student Name	Grade	Season	Semester
Criteria →				= NM	= 1
				= RE	= 1

Figure 16 – QBE query 4

Obtained Result/Answer 4

Course Designation	Student Name	Grade	Season	Semester
MÉTODOS QUANTITATIVOS	ANGELA RITA COSTA SALG.	13 NM		1
MÉTODOS QUANTITATIVOS	OLGA RAQUEL ARAÚJO MC	13 NM		1
MÉTODOS QUANTITATIVOS	JOSÉ MANUEL MOURÃO SE	8 NM		1
MÉTODOS QUANTITATIVOS	BRUNO JOSÉ NEVES DE SÁ	11 NM		1
MÉTODOS QUANTITATIVOS	JOSÉ MANUEL MOURÃO SE	7 RE		1
ORGANIZAÇÕES INTERNACIC	RICARDO LUÍS RANÇÃO AL	14 NM		1
ORGANIZAÇÕES INTERNACIC	MARCO JORGE MAMEDE M	16 NM		1
ORGANIZAÇÕES INTERNACIC	JOANA ANDREIA MAGALH	13 NM		1
ORGANIZAÇÕES INTERNACIC	JOANA PATRÍCIA LOURENÇ	11 NM		1
ORGANIZAÇÕES INTERNACIC	FATY ROSA SOARES SOUSA	1 NM		1
ORGANIZAÇÕES INTERNACIC	JOANA FILIPA FERNANDES	16 NM		1

Figure 17 – QBE query 4 result/answer

The simplicity and user-friendliness of QBE languages make them popular languages amongst business users of database systems, freeing them from the burden of the technical details. QBE languages are widely used across organizations' information systems and are important tools for business users accessing data stored in databases, helping in the decision making process.

5.2 Relational Calculus and Query-By-Example Languages

QBE languages are connected with relational calculus, which is a nonprocedural language. Relational calculus is based in a branch of mathematical logic called predicate calculus (Codd, 1971). Relational calculus allows the definition of database queries in a declarative way. There are two variations of relational calculus, namely, the tuple relational calculus, and the domain relational calculus. Both are formally equivalent (Date, 2004). Hereby we will use tuple relational calculus. In this context, a relational calculus query (Q) is a set of database tuples (t) satisfying some characteristics defined with a proposition (p):

$$Q = \{t/p(t)\}.$$

All QBE queries can be converted into relational calculus queries. Following, in Figure 18, are presented the relational calculus propositions corresponding to each of the QBE queries from Section 5.1.

Q1: {t.Teacher Name, t.Qualifications | TEACHER(t)}

Q2: {c.Course Designation, s.Student Name, r.Grade | COURSE(c) AND STUDENT(s) AND REGISTRY (r)}

Q3: {c.Course Designation, s.Student Name, r.Grade, e.Season, e.Semester | COURSE(c) AND STUDENT(s) AND REGISTRY (r) AND Season = "NM" AND Semester = 2}

Q4: {c.Course Designation, s.Student Name, r.Grade, e.Season, e.Semester | COURSE(c) AND STUDENT(s) AND REGISTRY (r) AND ((Season = "NM" AND Semester = 1) OR (Season = "RE" AND Semester = 1))}

Figure 18 – Relational calculus propositions for Q1 and Q2

PART III – RESEARCH APPROACH AND RESEARCH OUTPUTS

“Science is a convention, related to societal norms, expectations, values, etc. In its most conceptual sense, it is nothing more than the search for understanding. It would use whatever tools, techniques and approaches that are considered appropriate for the particular subject matter under study.” (Hirschheim, 1985 – pp 13). Each scientific discipline has its own appropriated and accepted approaches to research. Information systems (IS) is not an exception, and several approaches are being used and accepted in the field. In the research hereby presented, the approach was made using the framework of Design Science Research (DSR) that has a growing acceptance by the IS scientific community. The DSR for IS research is presented in Figure 19.

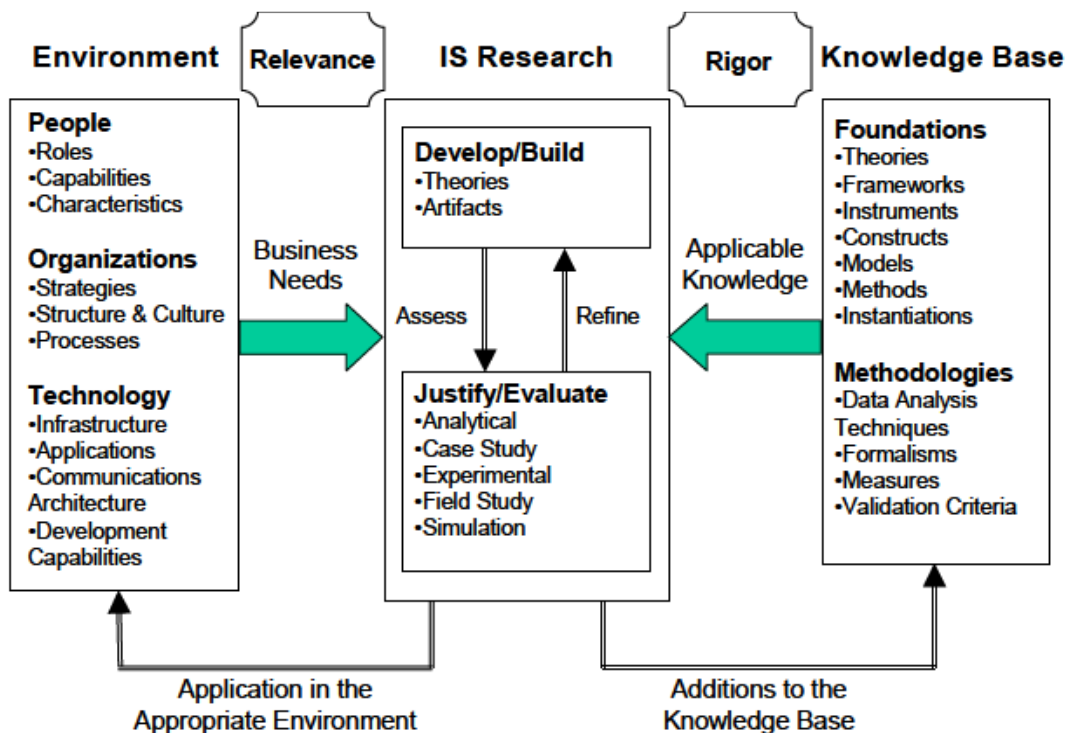


Figure 19 – Information Systems (IS) research framework (Hevner, 2004 – pp 80)

Here, in part III of this thesis, the research approach that has been adopted to conduct the research and the obtained research outputs are presented. It will start with the definition of the research problem, in chapter 6. Following, it is introduced the research framework that better

suits the stated research problem, in chapter 7. In chapter 8, the research methodology is presented. Chapter 9 brings in the concept of inductive data warehouse, presenting an example of an inductive data warehouse, in Section 9.1, and a generalization, in Section 9.2. Query-Models-By-Example language is introduced in chapter 10, presenting queries on data, in Section 10.1, queries on models, in Section 10.2, and queries on models and data, in Section 10.3, and relational calculus and QMBE, in Section 10.4. Part III concludes, in chapter 11, with the Query-Models-By-Example evaluation, doing a conceptual evaluation, in Section 11.1, and presenting the questionnaire to business users in Section 11.2, and concluding with some brief considerations, in section 11.3.

6 Research Problem

Organizations compete in environments whose complexity increases in a daily basis. Consequently, there are many demands that organizations must answer in time and adequately in order to survive and gain competitive advantage in those complex environments. In this context, computerized Decision Support Systems (DSS), in particular Business Intelligence (BI) systems, play an important role in order to improve decision making and thus conducting organizations' actions. BI systems are gaining momentum each day in organizations and have a fundamental role in these issues (Turban, Arosón, Liang & Sharda, 2007; Turban, Sharda, Arosón & King, 2008).

The usage of Data Mining (DM) tools in BI is increasing. The current notion of BI was coined by Gartner in 1989. KDD-1989, held in Chicago in August, 20th, is widely recognized as the first important event in the area of DM. This event has taken place regularly since then²³, and is nowadays the most important DM conference worldwide. Also Fayyad's book "Advances in knowledge discovery and data mining", published in 1996, was a landmark in the emergence of DM. Despite BI and DM having emerged roughly in the same epoch, they have different roots and as a consequence have significantly different characteristics (Kriegel, Borgwardt, Kröger, Pryakhin, Schubert & Zimek, 2007; Piatetsky-Shapiro, 2007). DM came up from scientific environments, thus it is not business oriented. DM tools still demand heavy work in order to obtain the intended results, hence needing the knowledge of DM specialists to explore its full potential value. On the contrary, BI is rooted in industry and business (Yermish, Miori, Yi, Malhotra & Klimberg, 2010), thus it is business oriented. As a result, BI tools are user-friendly and can easily be accessed and manipulated by business users.

From the literature review, presented in part II, it is evident that the majority of BI tools are directly manipulated by business users, allowing them to explore their potential value in a more effective way. The reason for this is related with the fact that BI tools are user-friendly, iterative, interactive, business oriented, and oriented to business activities. DM is an exception (McKnight, 2002; McKnight, 2003). Despite its usage in BI systems is increasing day by day, DM models are not directly manipulated by business users depending on reports from DM specialists. This way, business users could be unable to extract the potential business value contained in DM models. The complexity of DM models, as opposite to other BI tools, has been

²³ Source <http://www.sigkdd.org/conferences.php>, accessed 2/09/2011

identified as the key factor for this. From the literature review, it is also given evidence of the necessity to develop tools for DM that present the same characteristics of BI tools, namely being user-friendly, interactive, iterative, oriented to business users, and oriented to BI activities. Using DM tools possessing these characteristics, it will be possible that those tools can be directly manipulated by business users. This is aligned with the roots of DM and Knowledge Discovery in Databases (KDD) as stated in (Brachman & Anand, 1996), where KDD is presented as an iterative and interactive process, with many decisions being made by the user.

Quoting McKnight:

“much of data mining has been relegated to the domain of a special breed of expert, often holding a Ph.D. in statistics, mathematics or some scientific discipline. The mining process currently deployed in many organizations is not only time-consuming due to the challenge of the tools and the semantic gap between the front line and the statisticians; it is also non iterative in nature. Discovered nuggets flow from the miners to the front line and are only selectively interesting and actionable. (...) Mining tools that are interactive, visual, understandable, well- performing and work directly on the data warehouse/mart of the organization could be used by front-line workers for immediate and lasting business benefit.” (McKnight, 2002)

Realizing the importance of the aspects mentioned above, the recognition of this reality establishes the foundations for the research presented in this thesis. Binding DM to final business users of BI systems is considered a pertinent contribution. Accordingly, and based in the literature review presented in chapters 2 to 5, the research problem has been identified as:

Although final business users of BI systems directly access and manipulate data through the use of BI tools, thus being able of exploring the potential business value contained in databases (operational and data warehouses), they do not directly access and manipulate DM models and consequently the full potential business value hidden in DM models could be not completely explored. (Figure 20).

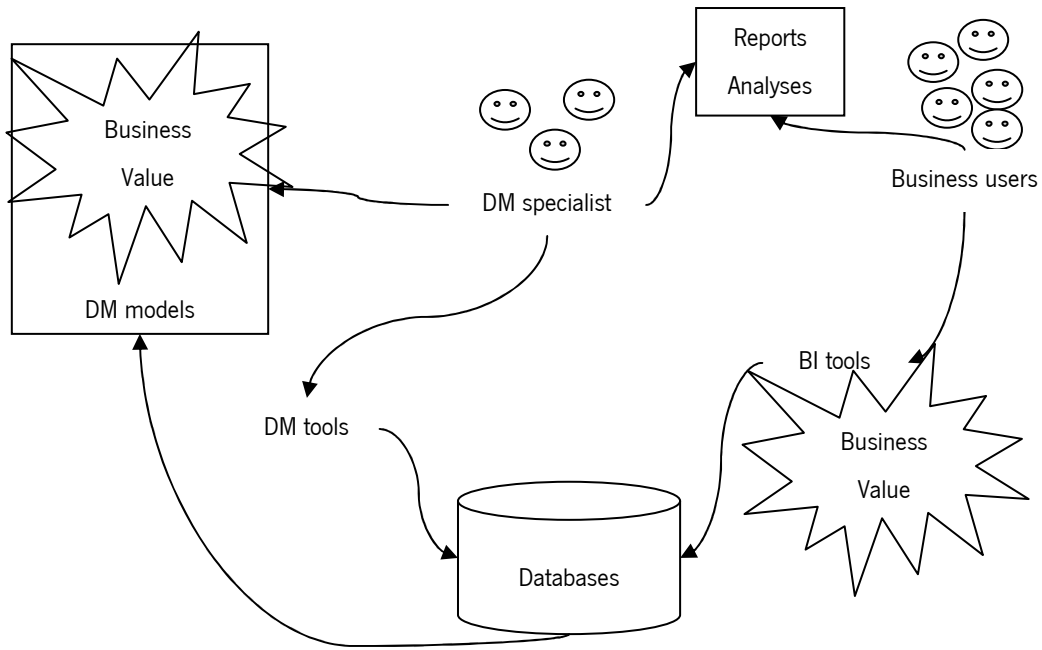


Figure 20 – Research Problem

The presented problem arises from the business needs existing in environments where BI systems include DM usage. The solution should achieve the following objectives:

- To allow business users to directly access and manipulate DM models;
- To be iterative;
- To be interactive;
- To work directly on DM models;
- To be easy to understand.

7 Research Framework

Design science research (DSR) was selected as the most appropriated to be used in the project described in this thesis. In (Hevner, March, Park & Ram, 2004; March & Smith, 1995) it is stated that DSR is rooted in the work of Simon (Simon, 1981) that presents the epistemological principles of the new sciences: The Sciences of the Artificial. DSR consists of two basic activities that are BUILD and EVALUATE, and its goal is UTILITY. The research is addressed through the building and evaluation of artifacts/outputs designed in order to meet identified business needs. The research problem or business needs are identified through the analyses of the considered environment, studying the interactions of the three IS components, namely people, organizations, and technologies, thus ensuring relevance. Artifacts are built based on the knowledge base (foundations and methodologies) of the field of study, and evaluated in order to assess, justify, and evaluate their adequacy. In the evaluation phase, two important questions concerning the artifact/output should be answered, namely, “Does it work?” and “Is it an improvement?”.

After many years of behavioral science domination in Information Systems (IS) research, design science is gaining popularity in the field (Carlsson, 2006; Hevner & Chatterje, 2010; Hevner et al, 2004; March & Smith, 1995; March & Storey, 2008; McKay & Marshall, 2005). In 2008 a special landmark occurred with the publication, in MIS Quarterly, of a special issue on DSR in the IS discipline. This special issue includes five important articles describing the application of DSR (Abbasi & Chen, 2008; Adomavicius, Bockstedt, Gupta & Kauffman, 2008; Lee, Wyner & Pentland, 2008; Parsons & Wand, 2008; Pries-Heje & Baskerville, 2008).

Based on the IS research framework presented in (Hevner et al, 2004), it is introduced in Figure 21 the research framework that supported the research presented in this thesis.

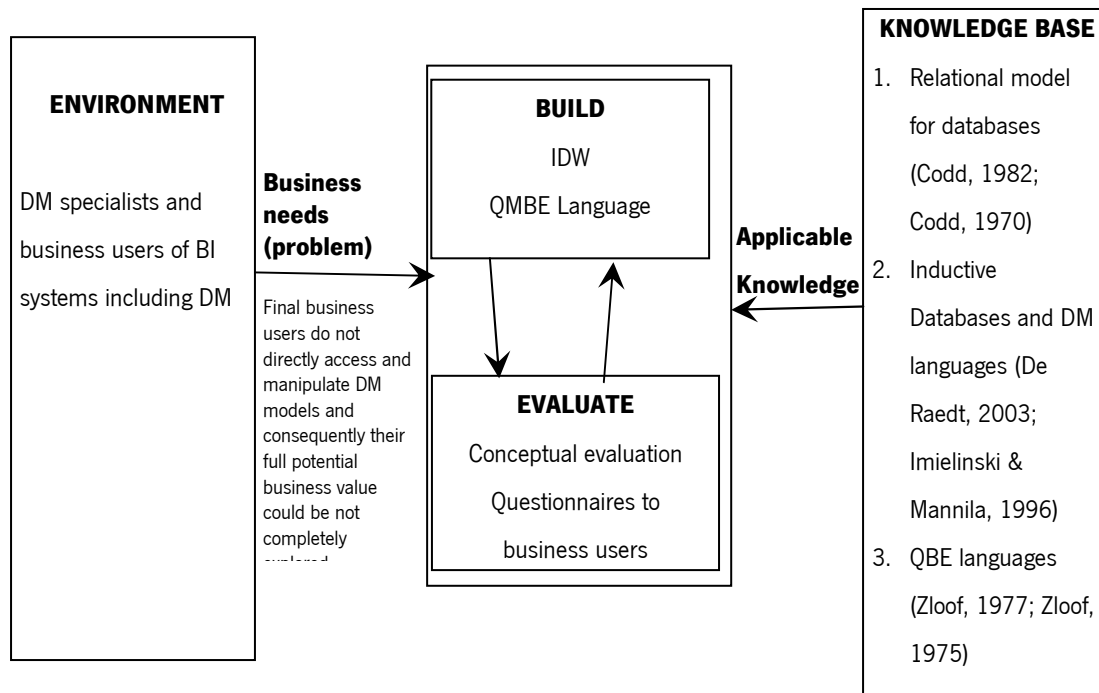


Figure 21 – Research Framework

The research problem, presented in chapter 7, was identified considering the constraints involved in an environment of DM specialists and business users of BI systems that include DM. The analysis of the knowledge base, formed by relational model for databases, inductive databases and DM languages, and Query-By-Example (QBE) languages, led to the selection of the Applicable Knowledge. The BUILD phase led to two outputs: Inductive Data Warehouse (IDW) concept and Query-Models-By-Example (QMBE) language specification. The EVALUATE phase was two-folded: a conceptual evaluation was formulated, and a questionnaire to business users was undertaken. The research methodology is presented in chapter 8.

The Design Science Research Methodology includes six steps/activities (Hevner & Chatterje, 2010 – pp 28-30):

- “ACTIVITY 1
Problem identification and motivation. Define the specific research problem and justify the value of a solution”;
- “ACTIVITY 2
Definition of the objectives for a solution. Infer the objectives of a solution from the problem definition and knowledge of what is possible and feasible”;
- “ACTIVITY 3

Design and development. Create the artifact”;

- “ACTIVITY 4

Demonstration. Demonstrate the use of the artifact to solve one or more instances of the problem”;

- “ACTIVITY 5

Evaluation. Observe and measure how well the artifact supports a solution to the problem”;

- “ACTIVITY 6

Communication. Communicate the problem and its importance, the artifact, its utility and novelty, the rigor of its design, and its effectiveness to researchers and other relevant audiences”.

All of these steps/activities were carried out in the research presented in this thesis. Activities 1 and 2 came up from the literature review, and are presented in chapter 6. Activity 3 is presented in chapter 8. Activity 4 is presented in chapters 9 and 10. Activity 5 is presented in chapters 11, 12 and 13. Activity 6 is achieved with this thesis and the publications listed in Appendix C.

8 Research Methodology

One of the main aspects of BI systems is that its user-friendly tools make systems truly available to final business users. As presented above, powerful analytical tools, such as DM, are still too complex and sophisticated for the average consumer of BI systems. In (McKnight, 2002) it is stated that bringing DM into the front line business personnel will increase their potential of attaining BI's high potential business value. Another fundamental issue that is pointed out as important 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. In this thesis it is considered that the framework of inductive databases, introduced in chapter 4 as a way to DM integration with relational databases, can also be a way to achieve the goal of a full integration of DM with BI, and leading to the end of, the already referred, DM isles in BI systems.

Taking these issues into consideration, as well as the research problem and the research framework, an architecture that allows an effective usage of DM by business users in BI systems, in order to conduct to DM integration with BI, was envisaged. This architecture should:

- bring DM into the front line business users;
- be iterative, visual, and understandable by front line business users;
- work directly on data.

It is considered that this can be achieved through a DM language that business users can understand and, consequently, use it to manipulate and query both DM models and data. Following these guidelines, an architecture for integration of DM with BI is presented in Figure 22, as an extension of the one that is presented in Figure 5, and intends to conduct to an effective usage of DM in BI. As far as we know, there is no similar architecture in the literature. This architecture implements the concept of Inductive Data Warehouses (IDW), which is a data warehouse storing data and data mining models at the same level, this is to say, both data and DM models are stored in the data warehouse tables and can be accessed and manipulated in the same way. It includes two additional modules: DM module and a new language named Query-Models-By-Example (QMBE).

The DM module extracts data from the Data Warehouse (DW), generates the DM models, and feeds the DW with DM models, storing them in tables of the DW. There is the possibility to

include as many models as needed by the user, and new models can be included just by adding new tables.

QMBE was developed and implemented as an extension of QBE languages, presented in chapter 5. Using QMBE the user is, thus, able to interact directly with the models, and to construct queries, including different criteria. Table 5 presents several business questions regarding DM models commonly posed by business users. All the business questions can be converted into queries to the system, defined in the QMBE language. The language has two important characteristics, which are interactivity, and iterativeness. These characteristics are inherited from QBE languages upon which QMBE is extended. The novelty of the QMBE language is that it is oriented to business users and to BI processes. This kind of approach allows business users to directly access and manipulate data and models, instead of relying in reports from DM specialists. This will bring DM to the front line business users, alike the other BI tools, thus allowing DM integration with BI.

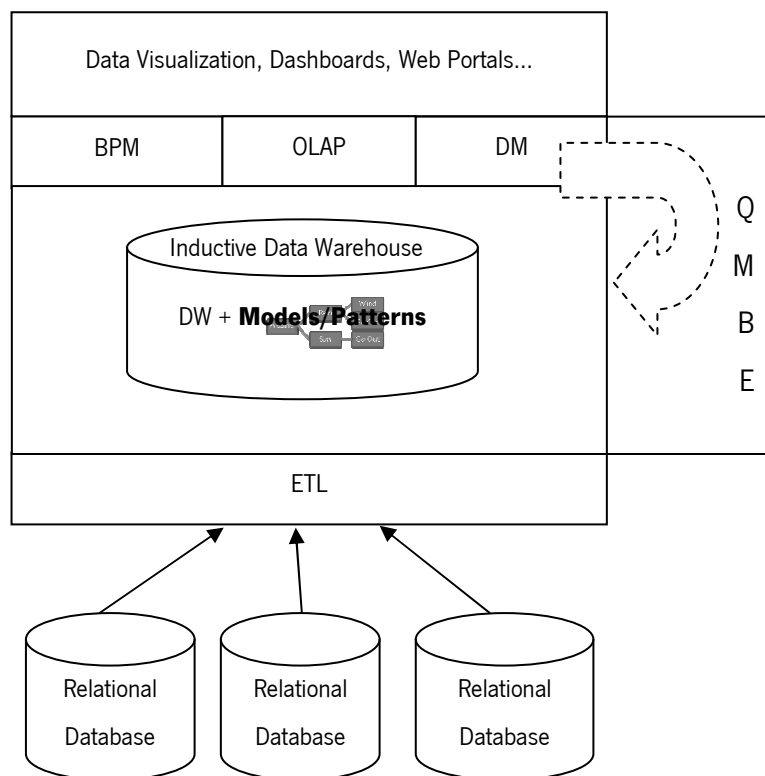


Figure 22 – Architecture for integration of Data Mining with Business Intelligence

Table 5 – Examples of business questions involving DM models

Queries on models	Queries on models and data
What are the characteristics of “good” students?	Select the potential students that can be “good” students.
What are the characteristics of “bad” students?	Select the potential students that can be “bad” students.
What are the characteristics of the students that do not conclude the grades according to initial schedule?	Select the potential students that cannot conclude the grades according to initial schedule.
Are there different types of students in the school?	

The project was developed using examples from a Portuguese Higher Education Institution (HEI). The presented architecture was implemented as a prototype. It was developed so that it is according to the inductive database framework, providing the features indicated in (Bonchi, Giannotti, Lucchesse, Orlando, Perego & Trasarti, 2007), and that were already indicated above in this thesis. Following, the indicated features and the respective way they were achieved in the prototype implementation are presented, namely:

- coupling with a database management system, since data and DM models are stored in the same database (the data warehouse) at the same level, namely in tables of the database;
- expressiveness of the query language, since business questions that can be posed by business users to the system can be converted into queries in the query language (QMBE);
- efficiency of the mining engine, which is guaranteed by the selection of the application used to generate the DM models; and
- the existence of a graphical user interface.

The DW was planned and implemented based on the multidimensional model presented in (Kimball & Ross, 2002). According to this model, two types of tables are considered in the DW: a fact table, and several dimension tables. The fact table is the main table in the dimensional model and stores measures of the business performance. The dimension tables stores the relevant business descriptors, and addresses how data will be analyzed. The business process, or fact, considered as adequate for the problem under study in this research, was students’ enrollment for examination. Several dimensions were also considered as adequate for the

problem under study: student, teacher, course, program, season, and level. Each one of these, fact and dimensions, gave rise to a table of the DW. Thus, the DW tables are: FACT TABLE, DIMENSION STUDENT TABLE, DIMENSION TEACHER TABLE, DIMENSION COURSE TABLE, DIMENSION PROGRAM TABLE, DIMENSION SEASON TABLE, AND DIMENSION LEVEL TABLE. The DW schema is presented in Figure 23.

FACT TABLE (<u>Season ID</u> , <u>Course ID</u> , <u>Program ID</u> , <u>Level ID</u> , <u>Student ID</u> , <u>Teacher ID</u> , Counter)
DIMENSION STUDENT TABLE (<u>Student ID</u> , Student name, Student Street, Student Zip code, Student Parish, Student Municipality, Student Area, Student Nationality, Student Gender, Student Age, Student qualification, Student admission type, Student admission level, Student secondary studies, Student secondary level (K12), Student secondary level (K11), Student secondary level (K10), Attendance type, Social scholarship?, Erasmus scholarship?)
DIMENSION TEACHER TABLE (<u>Teacher ID</u> , Teacher name, Teacher rank, Teacher qualification, Teacher years on duty, Teacher age)
DIMENSION COURSE TABLE (<u>Course ID</u> , Course designation, Department, #of theoretic hours, #of practical hours, #of theoretical-practical hours, Optional?, ECTS credits, Course semester, Course year)
DIMENSION PROGRAM TABLE (<u>Program ID</u> , Program designation)
DIMENSION SEASON TABLE (<u>Season ID</u> , Season description, Season Semester, Season Year)
DIMENSION LEVEL TABLE (<u>Level ID</u> , Level description)

Figure 23 – DW schema used in the research

Following, selected data was extracted from the DW. This data was processed and transformed in order to apply DM. DM was applied, using an open source DM tool, Weka⁴⁴, in order to obtain the DM model. At this point, classification rules were generated using an appropriate algorithm.

The IDW framework was used, thus the obtained DM models were stored on the DW in a specific table. More details are given in chapter 9.

⁴⁴ "Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Weka is open source software issued under the GNU General Public License.", in <http://www.cs.waikato.ac.nz/ml/weka/>, accessed 2/09/2011

QMBE language, described in chapter 10, was implemented and tested. The database management system (DBMS) used for the DW implementation was used to do that. The used DBMS contains the implementation of a QBE language that was extended in order to implement QMBE. The DBMS engine, including some additional programming code, managed the operationalization of the process.

The language was then evaluated. Firstly, a conceptual evaluation was done (Section 11.1). Next, the language was evaluated by business users who already used DM tools in BI systems, using a questionnaire. The questionnaire development and the analysis of the results are presented in Section 11.2.

9 Inductive Data Warehouse

“Inductive databases tightly integrate databases with data mining. The key ideas are that data and patterns (or models) are handled in the same way, and that an inductive query language allows the user to query and manipulate the patterns (or models) of interest” (De Raedt, 2003 – pp 69). Considering that DM models are stored in the database, thus stored in tables, they can be accessed and manipulated similarly to data. The inductive databases (IDB) framework was used in this research.

In the context of BI there can be said that an IDB contains both the DW and the knowledge base (KB), that is to say, the DM models. This way we can refer to this database as an Inductive Data Warehouse (IDW). Thus, an IDW is a DW that includes data and DM models, both stored in tables of the DW. This is an important concept in the realm of this research, since it focuses on making data mining available to business users. In an IDW data and DM models can be accessed by business users in the same way as data. The DM models are stored in the DW in specific tables: the model tables. It is possible to include several model tables, one for each generated model.

The presentation of the concept of IDW will be two folded. Firstly, the example presented in the previous section will be used. Next, a generalization will be made.

9.1 An example of an inductive Data Warehouse

In the presented example, the generated DM model corresponds to rules, since these were considered adequate for the problem under study. A rule is an IF-THEN expression of the form *IF antecedent THEN consequent*, written as:

$$\textit{antecedent} \Rightarrow \textit{consequent}$$

where antecedent and consequent are propositions of the form

$$V_1 \text{ cond}_1 C_1 \text{ AND } \dots \text{ AND } V_N \text{ cond}_N C_N$$

where V_1, \dots, V_N are variables; C_1, \dots, C_N are constants; and $\text{cond}_1, \dots, \text{cond}_N$ stands for < or > or = or <= or >=.

In the case of classifications rules, the consequent is of the form

$$V_i \text{ cond}_i C_i$$

where V_i is the target variable; C_i is a constant; and cond_i stands for < or > or = or <= or >=.

In this research, the classification rules were generated using Weka's class named as *weka.classifiers.rules.JRip*. "This class implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen as an optimized version of IREP."¹⁵ The algorithm is briefly described in appendix D.

At this point of the research the authors made some decisions. The first one was that only classification rules are addressed at that moment. This is due to the fact that the application domain focuses mainly on two DM tasks: classification and clustering. Classification can be addressed through classification rules, which can be easily understood by business users. Thus, classification rules were used in this research. Another decision was that the quality of the generated models was not a concern during this research, since this was not considered as an important issue to address at the moment.

The obtained DM model, included in appendix E, was then stored in a table of the database, named as MODEL TABLE. Each line of MODEL TABLE contains one rule. The schema for the model table considering rules was defined as follows: the first column corresponds to the rule identifier; the next two columns are rule confidence and support; the following column corresponds to the target DM variable; variables selected for data mining form the rest of the table columns, each variable corresponding to one column. The MODEL TABLE schema is presented in Figure 24. Joining the DW schema with the MODEL TABLE schema we get the IDW schema that is presented in Figure 25. Each rule is introduced into the MODEL TABLE as a line of the table. Data is introduced in a cell of the table whenever there is a constraint in the rule for the correspondent variable. Figure 26 contains examples of rules and their corresponding representation in the MODEL TABLE. This table, which stores DM models, is, in this context, similar to any other table belonging to the DW and thus DM models can be accessed and manipulated at the same level than data.

¹⁵ in <http://classes.engr.oregonstate.edu/eecs/winter2003/cs534/weka/weka-3-3-4/doc/weka.classifiers.rules.JRip.html>, accessed 2/09/2011

MODEL TABLE (Rule ID; Confidence; Support; Level description; Season ID; Season Semester; Season Year; Course ID; Department; #of theoretic hours; #of practical hours; #of theoretical-practical hours; Optional?; ECTS credits; Student Area; Student Nationality; Student Gender; Student Age; Student qualification; Student admission type; Attendance type; Program ID; Teacher rank; Teacher qualification; Teacher years on duty; Teacher age)

Figure 24 – Model Table schema used in the research

MODEL TABLE (Rule ID; Confidence; Support; Level description; Season ID; Season Semester; Season Year; Course ID; Department; #of theoretic hours; #of practical hours; #of theoretical-practical hours; Optional?; ECTS credits; Student Area; Student Nationality; Student Gender; Student Age; Student qualification; Student admission type; Attendance type; Program ID; Teacher rank; Teacher qualification; Teacher years on duty; Teacher age)

FACT TABLE (Season ID, Course ID, Program ID, Level ID, Student ID, Teacher ID, Counter)

DIMENSION STUDENT TABLE (Student ID, Student name, Student Street, Student Zip code, Student Parish, Student Municipality, Student Area, Student Nationality, Student Gender, Student Age, Student qualification, Student admission type, Student admission level, Student secondary studies, Student secondary level (K12), Student secondary level (K11), Student secondary level (K10), Attendance type, Social scholarship?, Erasmus scholarship?)

DIMENSION TEACHER TABLE (Teacher ID, Teacher name, Teacher rank, Teacher qualification, Teacher years on duty, Teacher age)

DIMENSION COURSE TABLE (Course ID, Course designation, Department, #of theoretic hours, #of practical hours, #of theoretical-practical hours, Optional?, ECTS credits, Course semester, Course year)

DIMENSION PROGRAM TABLE (Program ID, Program designation)

DIMENSION SEASON TABLE (Season ID, Season description, Season Semester, Season Year)

DIMENSION LEVEL TABLE (Level ID, Level description)

Figure 25 – IDW schema used in the research

Rule 8:
 (Program ID >= 3200) AND (Course ID >= 2218) AND (Course ID <= 2439) AND (Student Gender = F) AND (Teacher rank = Professor Adjunto) AND (ECTS credits >= 6) AND (Student Age >= 28) => Level description=Good

Rule 27:
 (Student Age >= 21) and (Department = Economia) and (ECTS credits <= 4) and (Teacher years on duty >= 22) and (Season ID = N) => Level description=Not present

Rule 32:
 (Season ID = R) and (Course ID <= 2355) and (Program ID <= 3100) and (Student Age <= 21) and (Student Area = BRAGA) => Level description=Not approved

Model Table:

Rule ID	Confidence	Support	Level description	Season ID	Season Semester	Season Year	Course ID	Department	
8	40,91%	66	Good				>=2218 and <=2439		...
27	44,04%	109	Not present	N				Economics	
32	31,18%	93	Not approved	R			<=2355		

#of theoretic hours	#of practical hours	#of theoretical-practical hours	Optional?	ECTS credits	Student Area	Student Nationality	Student Gender	Student Age	
				>=6			F	>=28	...
				<=4				>=21	
					BRAGA			<=21	

Student qualification	Student admission type	Attendance type	Program ID	Teacher rank	Teacher qualification	Teacher years on duty	Teacher age
			>=3200	Professor Adjunto			
						>=22	
			<=3100				

Figure 26 – Examples of rules and their corresponding representation in the Model Table

9.2 Generalization

Usually BI systems are supported by special databases, namely data warehouses (DW). For the sake of generality, consider a DW with one fact table named FACT_TABLE, and N dimension tables named DIMENSION_1, DIMENSION_2, DIMENSION_3, ..., DIMENSION_N. The fact table has one ID column, and N columns Dimension1, Dimension2, Dimension3, ..., DimensionN, each corresponding to one dimension table, and a column Fact. Each of the dimension tables has got several columns, each one corresponding to a variable that can be selected for DM. Consider for instance that DIMENSION_J has M_j variables, namely, ID_J, VarJ1, VarJ2, ..., VarJl, ..., VarJM_j.

In an IDW, DM models are stored in the database in one, or more, specific table, or tables. In this research only rules will be considered. Without losing generality, hereby only one table will

be considered and named MODEL_TABLE. The first column of the model table, ID, corresponds to the rule identifier. The next two columns, confidence and support, stand respectively for the rule confidence and for the rule support. The following column corresponds to the selected DM target variable that corresponds to one of the columns of one of the dimension tables. The L variables selected for data mining, each one corresponding to a column of one of the dimension tables included in the DW, form the rest of the table columns, namely, DMVar1, DMVar2, ..., DMVarL. Keep in mind that DMVar1, DMVar2, ... DMVarL of MODEL_TABLE are selected from all the columns of tables DIMENSION_1, or DIMENSION_2, ..., or DIMENSION_N. Thus, all the columns of the MODEL_TABLE are the same as some column of the dimension tables. In this manner MODEL TABLE is connected to the DW tables. The IDW general schema is presented in Figure 27.

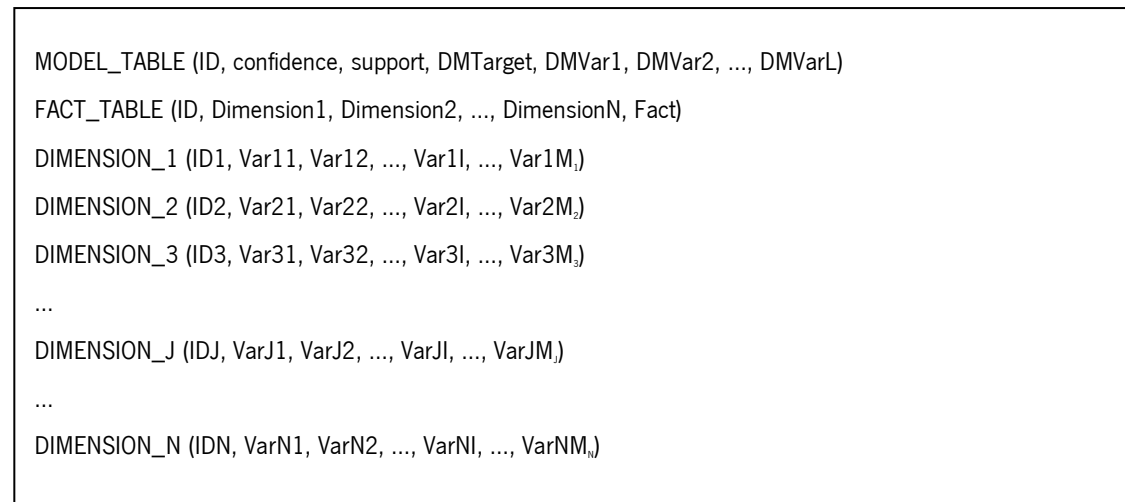


Figure 27 – IDW General Schema

Each rule is introduced in the MODEL_TABLE as a line of the table. Data is introduced in a cell of the table whenever there is a constraint in the rule for the correspondent variable, and is left blank (NULL) elsewhere. Consider, for instance, a general rule:

Rule I:

DMVar1 cond1 Value1 AND ... AND DMVarK condK ValueK AND ... => DMTarget condT ValueT; where cond1, ..., condK, condT stands for < or > or = or <= or >=.

Then the line (tuple) that corresponds to that rule is:

(I, valueC, ValueS, condT ValueT, cond1 Value1, ..., condK ValueK, ...).

New models can easily be added to the IDW by the simple introduction of model tables in the IDW, one for each model.

10 Query-Models-By-Example Language

In the research described in this thesis, a new language, named Query Models by Example (QMBE) was developed as an extension of QBE languages existing in some Relational Database Management Systems (RDBMS). Business users are able to interact directly with the models, and to construct queries as a way to obtain answers to ad-hoc business questions. Business questions can be converted into queries to the system, defined in the QMBE language. Like in RDBMS QBE languages, the user will be able to define different criteria, considered significant to business.

Business questions can be converted into queries in the QMBE language. To construct the query, the user will have to fill in a skeleton table (Figure 28).

Table →				
Column →				
Criteria →				

Figure 28 – Skeleton table for the QMBE language

The user will have to identify which are the tables, in the first line of the skeleton table; the corresponding columns that have the necessary data to answer the intended business question will have to be identified in the second line of the skeleton table. Specific criteria can be defined for each selected column, in the next lines of the skeleton table. More than one line can be considered for criteria. If criteria are defined in the same line, they are linked with AND. If criteria are defined in different lines, they are linked with OR.

In this section, QMBE language is introduced. There can be considered three types of QMBE queries (Figure 29), namely:

- queries on data, corresponding to traditional QBE languages;
- queries on models, corresponding to QMBE extensions; and
- queries on models and data, corresponding also to QMBE extensions.

In all these three cases, examples of business questions will be presented based on the IDW schema from Figure 25. There will be also presented the correspondent queries in QMBE, as well as the answers received from the system, in response to the queries. General cases, based on the general IDW schema presented in Figure 27, will also be presented. The chapter

will conclude with the presentation of the relational calculus sentences that correspond to each one of those QMBE queries.

QMBE

Traditional QBE: queries on data	Extension 1: Queries on models	Extension 2: Queries on models and data
-------------------------------------	-----------------------------------	---

Figure 29 – Types of QMBE queries

10.1 Queries on data

This type of queries corresponds to traditional QBE languages. Using this type of queries, business users are able to manipulate data stored in databases. The MODEL TABLE is not involved in this type of queries. Examples of queries using the IDW schema of Figure 25 are presents in section 10.1.1, and a general query using the IDW general schema of Figure 27 is presents in section 10.1.2.

10.1.1 Using an example

The presented business questions were considered having in mind that different types of queries were involved. Query 1 involves data from only one table in the IDW. Query 2 involves data from more than one table in the IDW. Query 3 involves only one criterion. Query 4 involves more than one criterion.

Business Question 1

Where are students from?

QMBE query 1

In this example, there is one table containing the necessary data, namely DIMENSION STUDENT table. The necessary columns are: column *Student Name*, and column *Student Zip code* from *Dimension Student table*. There are no criteria to consider. The query is presented in Figure 30 and the answer to this query is presented in *Figure 31*.

Table →	DIMENSION STUDENT	DIMENSION STUDENT		
Column →	Student Name	Student Zip code		
Criteria →				

Figure 30 – QMBE Query 1

Answer 1

Student name	Student Zip code
DANA GODINHO LEÃO SEQUEIRA LINHAS	RIO TINTO
DENIZ ERIZ	S. MAMEDE DE INFESTA
ARZU AH	S. MAMEDE DE INFESTA
DINU ANA-MARIA	RIO TINTO
TINE COUMANS	RIO TINTO
JIRI STOKLASA	PORTO
IRENA TAUEROVÁ	PORTO
ZDENEK ROTH	PORTO
MEHMET SOYLU	S.MAMEDE DE INFESTA

Figure 31 – Answer to query 1

Business Question 2

Which are students' classifications (levels) in each course?

QMBE query 2

In this example, tables and corresponding columns containing the necessary data are: column *Student Name* from DIMENSION STUDENT table, column *Course ID* from DIMENSION COURSE table, and column *Level description* from DIMENSION LEVEL table. There are no criteria to consider. The query is presented in Figure 32 and the answer to this query is presented in Figure 33.

Table →	DIMENSION STUDENT	DIMENSION COURSE	DIMENSION LEVEL	
Column →	Student Name	Course ID	Level description	
Criteria →				

Figure 32 – QMBE Query 2

Answer 2

JOANA PATRÍCIA LOURENÇO DIAS	2413	Satisfactory
JOANA PATRÍCIA LOURENÇO DIAS	2413	Good
FATY ROSA SOARES SOUSA PINTO	2413	Fair
FATY ROSA SOARES SOUSA PINTO	2413	Good
JOANA ISABEL OLIVEIRA MORAIS DE ALMEIDA	2413	Very Good
JOANA ISABEL OLIVEIRA MORAIS DE ALMEIDA	2413	Very Good
TIAGO ANDRÉ NOGUEIRA DA CRUZ	2413	Good
TIAGO ANDRÉ NOGUEIRA DA CRUZ	2413	Good
FATY ROSA SOARES SOUSA PINTO	2413	Satisfactory
JOANA PATRÍCIA LOURENÇO DIAS	2413	Good

Figure 33 – Answer to query 2

Business Question 3

Who are the students with the best classification (Level Very Good) in each course?

QMBE query 3

In this example, tables and corresponding columns containing the necessary data are the same as the ones in example 2. There is one criterion to consider: *Level description*="Very Good". The query is presented in Figure 34 and the answer to this query is presented in Figure 35.

Table →	DIMENSION STUDENT	DIMENSION COURSE	DIMENSION LEVEL	
Column →	Student Name	Course ID	Level description	
Criteria →			"Very Good"	

Figure 34 – QMBE Query 3

Answer 3

Student name	Course ID	Level description
JOANA ISABEL OLIVEIRA MORAIS DE ALMEIDA	2413	Very Good
JOANA ISABEL OLIVEIRA MORAIS DE ALMEIDA	2413	Very Good
PAULO SÉRGIO DA COSTA MOREIRA	2358	Very Good
JOÃO PEDRO TAVARES DE SOUSA	2352	Very Good
ANA LAURA DA SILVA FERREIRA	2021	Very Good
CRISTIANA SOFIA DIAS PINHEIRO	2021	Very Good
DAMIÃO JOSÉ ANDRÉS RODRIGUES	2021	Very Good
SARA CARVALHO TEIXEIRA	2021	Very Good
DANIELA DA SILVA FERREIRA	2085	Very Good

Figure 35 – Answer to query 3

Business Question 4

Which are the names of Not Approved students in Mathematics or Statistics courses?

QMBE query 4

In this example, tables and corresponding columns containing the necessary data are: column *Level description* from DIMENSION LEVEL table, column *Student Name* from DIMENSION STUDENT table, and column *Course Name* from DIMENSION COURSE table. Criteria must be defined for column *Level description* and for column *Course Name*: (*Level description*="Not Approved" AND *Course name*="Mathematics") OR (*Level description*="Not Approved" AND *Course name*="Statistics"). The query is presented in Figure 36 and the answer to this query is presented in Figure 37.

Table →	DIMENSION LEVEL	DIMENSION STUDENT	DIMENSION COURSE	
Column →	Level description	Student Name	Course name	
Criteria →	"Not Approved"		"Mathematics"	
Criteria →	"Not Approved"		"Statistics"	

Figure 36 – QMBE Query 4

Answer 4

Level description	Student name	Course designation
Not Approved	ANA PAULA PEREIRA DA SILVA NEVES	Mathematics
Not Approved	ANA PAULA PEREIRA DA SILVA NEVES	Mathematics
Not Approved	ANA PAULA SILVA	Mathematics
Not Approved	ANA RAQUEL FERREIRA RAIMUNDO	Statistics
Not Approved	ANA RAQUEL GONCALVES SIMAO	Mathematics
Not Approved	ANA RITA DE CASTRO FERREIRA	Mathematics
Not Approved	ANA RITA DE CASTRO FERREIRA	Mathematics
Not Approved	ANA RITA TEIXEIRA CARDOSO	Mathematics
Not Approved	ANA SOFIA GONÇALVES PINTO	Mathematics
Not Approved	ANA SOFIA GONÇALVES PINTO	Mathematics
Not Approved	ANA SOFIA MATOS DA SILVA	Statistics
Not Approved	ANABELA VIEIRA NUNES	Mathematics
Not Approved	ANDRE EMANUEL GOMES QUEIROS	Mathematics
Not Approved	ANDRE EMANUEL GOMES QUEIROS	Mathematics

Figure 37 – Answer to query 4

10.1.2 A general query

Generally speaking, queries on data involve columns from any of the tables of the IDW, except the MODEL_TABLE, for instance DIMENSION_J and FACT_TABLE. Similarly, criteria can be defined for any column.

Business Question I

What are the data from Dimension J table that corresponds to Fact (of FACT TABLE) equal to a certain value (value)?

QMBE query I

The query is presented in Figure 38.

Table →	FACT_TABLE	DIMENSION_J	...	DIMENSION_J
Column →	Fact	Var11	...	Var1N
Criteria →	value			

Figure 38 – QMBE query I

10.2 Queries on models

This type of queries corresponds to an extension of traditional QBE languages. With these queries, business users are able to manipulate DM models stored on database tables. Example of queries using the IDW schema of Figure 25 are presented in section 10.2.1, and a general query using the IDW general schema of Figure 27 is presents in section 10.2.2.

10.2.1 Using an example

Clearly, only the MODEL TABLE is involved in this type of queries. The obtained answers are rules. Query 5 and query 7 involve all the columns from MODEL TABLE. Query 6 involves some of the columns from MODEL TABLE. Query 5 and Query 6 criterion involves only the DM target variable. Query 7 criterion involves other variable(s) than the DM target variable.

Business Question 5

What are the characteristics of Good classifications (levels)?

QMBE query 5

Since the user needs to know all the characteristics, all the MODEL TABLE columns are necessary. Criteria must be defined for column *Level description*. *Level description* = "Good". The query is presented in Figure 39 and the answer to this query is presented in Figure 40.

Table →	MODEL TABLE	MODEL TABLE	...	MODEL TABLE
Column →	Level description	Course Name	...	Student's Age
Criteria →	"Good"			

Figure 39 – QMBE Query 5

Answer 5

Role ID	Level description	Seat	Seac	Season	Course ID	Department	#of the o	#of on	#of the	Optional?	ECTS	Student Area	Student f	Stude	Student Age
8	Good				>=2218 and <=						>=6		F		>=28
9	Good				>=2222								F		>=33
10	Good	N			>=2218						<=3		F		
11	Good	N			>=2222	Geosic							F		>=26

Figure 40 – Answer to query 5

Business Question 6

What are the characteristics of Not Approved students?

QMBE query 6

Since the user needs to know only students' characteristics, only the MODEL TABLE's columns concerning students' characteristics will be considered. Criteria must be defined for column *Level description*: *Level description="Not Approved"*. The query is presented in Figure 41 and the answer to this query is presented in Figure 42.

Table →	MODEL TABLE	...	MODEL TABLE	...	MODEL TABLE
Column →	Level description	...	Student Gender	...	Student's Age
Criteria →	"Not Approved"				

Figure 41 – QMBE Query 6

Answer 6

Level description	Student Area	Student Nationality	Student Gender	Student Age	Student qualification	Student admission type	Attendance type
Not Approved							
Not Approved							
Not Approved	BRAGA			<=21			
Not Approved				<=28			T
Not Approved				<=28			T

Figure 42 – Answer to query 6

Business Question 7

Which are the rules concerning female students?

QMBE query 7

Since the user needs to know all the characteristics, all the *MODEL TABLE*'s columns are necessary. Criteria must be defined for column *Student Gender: Student Gender="F"*. The query is presented in Figure 43 and the answer to this query is presented in Figure 44.

Table →	MODEL TABLE	...	MODEL TABLE	...	MODEL TABLE
Column →	Level description	...	Student Gender	...	Student's Age
Criteria →			"F"		

Figure 43 – QMBE Query 7

Answer 7

Rule ID	Level description	Seas	Sease	Season	Course ID	Department	#of theo	#of pr	#of the	Optional?	ECTS	Student Area	Student I	Stude	Student Age
5	Não Inscrito	N			>=2434									F	>=35
8	Good				>=2218 and <=						>=6			F	>=28
9	Good				>=2222									F	>=33
10	Good	N			>=2218						<=3			F	
11	Good	N			>=2222	Gestão								F	>=26

Figure 44 – Answer to query 7

10.2.2 A general query

Generally speaking, queries on models may involve any of the columns of the *MODEL_TABLE* and criteria can be defined for any column.

Business Question J

What are the rules of *MODEL TABLE* that correspond to DM Target equal to a certain value (value)?

QMBE query J

The query is presented in Figure 45.

Table →	MODEL_TABLE	MODEL_TABLE	...	MODEL_TABLE
Column →	DMTarget	DMVar1	...	DMVarL
Criteria →	value			

Figure 45 – QMBE query J

10.3 Queries on models and data

This type of queries corresponds to an extension of traditional QBE languages. By using these queries, business users are able to manipulate both data and models stored in databases. All the IDW tables are involved in this type of queries. Examples of queries using the IDW schema of Figure 25 are presented in section 10.3.1, and a general query using the IDW general schema of Figure 27 is presented in section 10.3.2.

10.3.1 Using an example

After analyzing models by direct manipulation, the business user can be interested in selecting the data that corresponds to the relevant model's rules. For instance, the business user can be interested in selecting new students who have the characteristics of "bad students" (Not Approved levels) according to some rule(s), in order to develop a special program to improve new students' classifications (levels). This can be done through the use of both data tables and model tables.

Consider for instance that, through direct manipulation of the rules stored in MODEL TABLE by business users, "bad students" have been identified with the characteristics defined by rule 32 (Figure 26).

Business Question 8

Who are the new students that correspond to the characteristics defined by rule 32, that is to say, that are potentially "bad students"?

QMBE query 8

In this example, all the columns from DIMENSION STUDENT table are included. Also included are the columns, and correspondent tables, that have criteria defined by rule 32 namely, Course ID from DIMENSION COURSE table, Student area, and Student age from DIMENSION STUDENT table, and Program ID from DIMENSION PROGRAM table. Data from the MODEL TABLE that correspond to criteria are passed to the skeleton table in the third line and in the corresponding column (Figure 46). Criteria must also be defined for column Enrollment Date from DIMENSION STUDENT table, in order to select only new students: Enrollment Date \geq season's start date. The query is presented in Figure 46 and the answer to this query is presented in Figure 47.

Rule ID	Confidence	Support	Level description	Season ID	Season Semester	Season Year	Course ID	Department
32	31,18%	93	Not Approved	R			<=2355	

#of theoretic hours	#of practical hours	#of theoretical-practical hours	Optional?	ECTS credits	Student Area	Student Nationality	Student Gender	Student Age
					BRAGA			<=21

Student qualification	Student admission type	Attendance type	Program ID	Teacher rank	Teacher qualification	Teacher years on duty	Teacher age
			<=3100				

Table →	DIMENSION STUDENT	...	DIMENSION STUDENT	...	DIMENSION STUDENT	...	DIMENSION STUDENT	DIMENSION COURSE	DIMENSION PROGRAM
Column →	Student Id	...	Student Area	...	Student Age	...	Student Enrolment Date	Course ID	Program ID
Criteria →			BRAGA		<=21		>= Season's start date	<=2355	<=3100

Figure 46 – QMBE Query 8

Answer 8

Student ID	Student Municipality	Program ID	Course ID
2020419	GUARDA	3100	1132
2020419	GUARDA	3000	1133
2020419	GUARDA	3100	1133
2020419	GUARDA	3000	1135
2020419	GUARDA	3100	1136
2020419	GUARDA	3000	1160
2020419	GUARDA	3100	1160

Figure 47 – Answer to query 8

10.3.2 A general query

Queries on models and data may involve columns from all the tables of the IDW and criteria can be defined for any column.

Business Question K

What are the data from DIMENSION J that corresponds to a pre-selected rule from MODEL TABLE, for instance, rule I (section 9.2)?

QMBE query K:

The query is presented in Figure 48.

Table →	DIMENSION_J	...	DIMENSION_J	...	DIMENSION_J	...
Column →	VarJ1		VarJ ₁		VarJ _k	
Criteria →			cond1 Value1		condK ValueK	

Figure 48 – QMBE query K

10.4 Relational calculus and QMBE

QBE languages, presented in chapter 5, are connected to relational calculus and so is QMBE. Just like for traditional QBE queries, all QMBE queries can be written as relational calculus queries. In the same way as for traditional QBE, if QMBE is considered, a relational calculus query (Q) is a set of database tuples (t) satisfying some characteristics defined with a proposition (p):

$$Q: \{t/p(t)\}.$$

This is also valid for both types of QMBE extensions to QBE.

Following, relational calculus queries corresponding to each of the QMBE queries previously introduced, are presented, queries on data in section 10.4.1, queries on models in section 10.4.2, and queries on models and data on section 10.4.3.

10.4.1 Traditional QBE: Queries on data

Following, are presented the relational calculus propositions corresponding to each of the QMBE queries from Section 10.1, which corresponds to this type of QMBE queries.

Relational Calculus Query 1

$$Q1 = \{s.Student Name, s.Student Zip Code \mid Dimension Student(s)\}$$

Relational Calculus Query 2

$$Q2 = \{s.Student Name, c.Couse ID, l.Level description \mid Dimension Student(s) \text{ AND } Dimension Course(c) \text{ AND } Dimension Level(l)\}$$

Relational Calculus Query 3

$$Q3 = \{s.Student\ Name, c.Course\ ID, l.Level\ description \mid Dimension\ Student(s)\ AND \\ Dimension\ Course(c)\ AND\ Dimension\ Level(l)\ AND\ l.Level\ description="Very\ Good"\}$$

Relational Calculus Query 4

$$Q4 = \{l.Level\ description, s.Student\ Name, c.Course\ ID \mid Dimension\ Level(l)\ AND \\ Dimension\ Student(s)\ AND\ Dimension\ Course(c)\ AND\ (l.Level\ description="Not\ Approved" \\ AND\ c.Course\ name="Mathematics")\ OR\ (l.Level\ description="Not\ Approved" \\ AND\ c.Course\ name="Statistics")\}$$

Relational Calculus Query I:

$$QI = \{f.Fact, d \mid FACT\ TABLE(f)\ AND\ DIMENSION_J(d)\ AND\ f.Fact = value\}$$

10.4.2 QMBE Extension 1: queries on models

Considering this type of QMBE extension, all QMBE queries can be converted to relational calculus because all DM models are stored in a database table. Considering rules, which are applied in this research, each rule is a database tuple since it is stored as a line in a database table. Following, are the relational calculus propositions corresponding to each of the QMBE queries from Section 10.2 that corresponds to this type of extension.

Relational Calculus Query 5

$$Q5 = \{m \mid Model\ Table(m)\ AND\ m.Level\ description="Good"\}$$

Relational Calculus Query 6

$$Q6 = \{m.Level\ description, m.Student\ Area, m.Student\ Nationality, m.Student\ Gender, \\ m.Student\ Age, m.Student\ qualification, m.Student\ admission\ type, m.Attendance \\ type \mid Model\ Table(m)\ AND\ m.Level\ description="Not\ Approved"\}$$

Relational Calculus Query 7

$$Q7 = \{m \mid Model\ Table(m)\ AND\ m.Student\ Gender="F"\}$$

Relational Calculus Query J:

$$QJ = \{m \mid MODEL\ TABLE(m)\ AND\ m.DMTarget=value\}$$

10.4.3 QMBE Extension 2: queries on models and data

Considering this type of QMBE extension, all QMBE queries can also be converted to relational calculus because all DM models are stored in a database table at the same level than data. Following, are the relational calculus propositions corresponding to each of the QMBE queries from Section 10.3. that corresponds to this type of extension.

Relational Calculus Query 8

Q8: {s, c.Course ID, p.Program ID | Dimension Student(s) Dimension Course(c) AND Dimension Program(p) AND Model Table(m) AND s.Student Area=m.Student Area AND s.Student Age=m.Student Age AND s.Student Enrolment Date>=Seasons' Start Date AND c.Course ID = m.Course ID AND p.Program ID=m.Program ID}

Relational calculus Query K:

QK: {dJ | DIMENSION_J(dJ) AND VarJ1 cond1 value1 AND ... AND VarJK condK ValueK AND ...}

11 Query-Models-By-Example Evaluation

Design science consists of two basic activities that are BUILD and EVALUATE, and its goal is utility. Artifacts are built based on the knowledge base (foundations and methodologies) of the field of study, and evaluated in order to assess, justify, and evaluate its adequacy. As stated in chapter 7, two important questions must be answered: “Does it work?” and “Is it an improvement?”

Hereby, a two-folded evaluation was undertaken, beginning with a conceptual evaluation and following with a questionnaire to business users.

11.1 Conceptual evaluation

The architecture presented in Figure 22 is based in two important frameworks: inductive databases (De Raedt, 2003; Imielinski & Mannila, 1996) and query-by-example languages (Zloof, 1975; Zloof, 1977). These two, constitute the fundamental kernel of the knowledge base that supports this research.

The IDB framework was used because, according to this framework, DM models are stored in databases at the same level than data. This allows DM models access to be made in a way that is similar to data access. Therefore, users can directly access and manipulate DM models as requested. The IDB framework was adapted to the context of BI through the introduction of the concept of Inductive Data Warehouse (IDW). An IDW includes data and DM models both stored in database tables.

It is also important to consider the inductive language, since it allows the users to access data and DM models stored in the IDW. The philosophy of Query-By-Example (QBE) languages was used, since those languages present the desired characteristics, namely being user-friendly, interactive, iterative, and oriented to business users. A new language named QMBE was developed based in this philosophy, thus having the same characteristics. Since this language is developed in the context of BI, it is also oriented to BI activities.

The architecture, including IDW and QMBE language, was implemented as a prototype in the considered environment and used in different and controlled situations, proving that the concepts are viable and can be applied. So the answer to the first question, referred in chapter 7, “Does it work?” is “Yes, it works.” In addition, as a consequence of the foundations, the proposed solution accomplishes the objectives that were defined for it in chapter 6, namely:

- To allow business users to directly access and manipulate DM models;
- To be iterative;
- To be interactive;
- To work directly on DM models;
- To be easy to understand.

11.2 Questionnaire to business users

QMBE was developed to reach, mainly, business users of BI systems incorporating DM, that is to say, individuals using BI systems and DM in a BI context but who are not necessarily DM specialists. In order to evaluate QMBE language from the perspective of these potential users, an online questionnaire was created. The questionnaire is included in Appendix A. Following the study premises are presented. The questionnaire structure is presented in section 11.2.1, and the analysis of the questionnaire responses is presented in section 11.2.2.

11.2.1 Questionnaire structure

The questionnaire was developed intending to achieve the following goals:

- To determine respondents' experience using BI and DM;
- To determine respondents' sort of usage of DM;
- To determine the degree of importance respondents assign to DM usage in the support to decision making;
- To obtain respondents' general perspective about the language;
- To ascertain whether QMBE offers advantages for business users analyzing DM models;
- To ascertain whether QMBE achieves the intended features (user-friendliness, iterativeness and interactivity, oriented to business users, oriented to BI activities, brings benefits to decision making).

In order to achieve these goals, the questionnaire was structured with the definition of three parts:

- Part 1 intends to check respondents' experience using BI systems and DM. It also intends to determine respondents' sort of usage of DM and the degree of importance they assign to DM usage in the support to decision making. This part of the questionnaire includes four questions. The first two questions are closed-ended

questions, and intend to check out respondents' experience on using BI systems and DM. Question 3 is a multiple choice question based in a 5-point Likert scale, ranging from 1 (Not at all important) to 5 (Very important), and asks about the importance respondents assign to DM. Question 4 is an open-ended question that asks about how the respondents use DM in their own organizations.

- Part 2 includes a movie¹⁶ presenting a tutorial of the QMBE language, in order to present the language to the respondents and let them get aware of the main QMBE characteristics. The movie tries to explain how the language works in practice and thus it does not present technical characteristics of the language. This part does not include any questions.
- Part 3 tries to obtain respondents' opinions about QMBE. A twofold approach was made. On one hand, it is proposed to obtain respondents' general perspective about the language. On the other hand, it was intended to ascertain whether the respondents consider that QMBE offers advantages for business users in analyzing DM models. This part of the questionnaire is composed of five questions. Questions 1 and 5 are open-ended questions, intended to collect respondents' personal opinions about the language. Question 2 encompasses eight statements about which respondents must express their opinion considering two hypotheses: DM models and DM models with QMBE. Each case is based in 5-point Likert scale, ranging from 1 (Strongly disagree) to 5 (Strongly agree). Question 3 intends to determine the respondents' will to adopt the presented language and is based in a 5-point Likert scale from 1 (Not at all) to 5 (Certainly). Part three of the questionnaire ends by asking the respondents to leave their email address if interested in knowing the results of this study.

Table 6 includes the list of the questionnaire goals and the questions developed in order to achieve those goals.

In order to find respondents to this survey, emails were sent to several mailing lists of international associations, and to other contacts.

¹⁶ The movie can be accessed at

<https://skydrive.live.com/?cid=48ca25fbc15d7192&sc=photos&ref=2&id=48CA25FBC15D7192%21119&sf=1>

Table 6 – Questionnaire goals and associated questions

Goal	Questions
To determine respondents' experience using BI and DM	1.1; 1.2 ; 1.4
To determine respondents' sort of usage of DM	1.4
To determine the degree of importance respondents assign to DM usage in the support to decision making	1.3
To obtain respondents' general perspective about the language	3.1; 3.4
To ascertain whether QMBE offers advantages for business users analyzing DM models	3.2
To ascertain whether QMBE achieves the intended features (user-friendliness, iterativeness and interactivity, oriented to business users, oriented to BI activities, brings benefits to decision making)	3.2

11.2.2 Analysis of the questionnaire results

With this questionnaire 16 valid responses were obtained. Characterization of the respondents is presented in section 11.2.2.1, and respondents opinion about QMBE is presented in section 11.2.2.2.

11.2.2.1 Characterization of the respondents

Part 1 - Question 1 and Question 2

Table 7 presents a summary of the answers to the first two questions in part 1 of the questionnaire, namely, "how long have respondents been using BI" and "how long have respondents been using DM". The means are about 5 and 4 years, respectively. Thirteen of the respondents have got three or more years using BI to support decision making, and eleven of the respondents have got three or more years using DM to support decision making. Thus, it can be considered that the respondents have got enough experience in using BI and DM.

	How long using BI (in years)	How long using DM (in years)
Mean	5,24	4,39
Standard deviation	3,15	3,21
Maximum	15	10
Minimum ¹⁷	0,08	0,08

Table 7 – Respondents' experience using BI and using DM

Part 1 - Question 3

Respondents consider that the use of DM to support decision making is Important/Very important (question 3, part 1), as can be concluded from the analysis of the graph presented in Figure 49. It is important to notice that the only respondent, who answered “1”, had been using BI and DM from only a month.

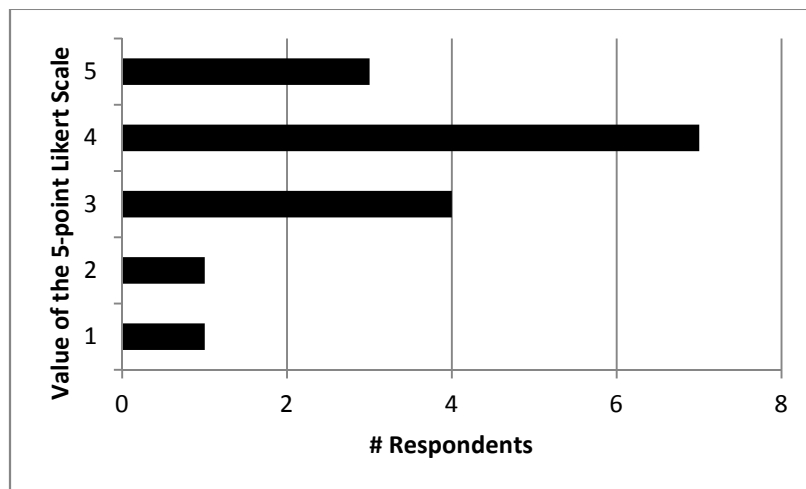


Figure 49 – Respondents' experience about the importance of DM (Graph)

Part 1 - Question 4

Only one of the respondents has not answered question 4 from part 1 of the questionnaire. The other respondents were a little vague about how they are using DM. Despite that, it can be

¹⁷ These values corresponds to 1 month usage

concluded that all of them use DM to help decision making. Thus it can be considered that respondents are more likely to be business users.

The answers to the questions belonging to the first part of the questionnaire allow to conclude that the respondents match the intended characteristics for the respondents: business users already using DM in BI systems.

11.2.2.2 Respondents opinion about QMBE

Part 3 - Question 1

The general opinion of the respondents about QMBE was positive (question 1, part 3). Only two of the respondents do not answer the question. Most of the respondents employ expressions like “Excellent”; “Good”, “Effective”, “Positive”, and “Very good”. One of the respondents uses “Normal”, and other uses the expression “Not too impressed”. It is also considered by the respondents that QMBE is “easy to understand” and is “useful to end-users” and “non-computer programmers”. There are some answers worthy of being highlighted, which are presented in Table 8.

Table 8 – Respondents’ general reaction to QMBE

“Seems like an intuitive to use interface that could be used to produce results quickly”
“A new approach to DM is highly desired.”
“Not too impressed ...but it looks easy to understand - can perhaps be done by many end-users.”
“It is very good. It seems to be easier to use for non-computer programmers.”

Part 3 - Question 2

Table 9 summarizes the answers to question 2 of part 3 of the questionnaire. Analyzing the results it is verified that, for all the statements, means are higher in the case of “DM with QMBE” thus it can be concluded that with QMBE, DM models:

- are easier to understand,
- are easier to use in practice,
- are more oriented to business users,
- are more oriented to BI activities,
- full potential can be better explored,
- better help decision making,
- bring more benefits to higher education institutions,
- bring more benefits to organizations in general.

Hypothesis tests were performed for all the statements in order to check out if these differences are statistically significant. The tests are included in Appendix B.

For each question, the null hypothesis (H_0) was tested against the alternative hypothesis (H_1). Each one of the respondents were asked to choose the adequate answer in a 5-point Likert scale considering for two situations, namely, *DM models* and *DM models with QMBE*, thus generating two paired samples. The null hypothesis considers that there are no significant differences between the means of the two samples, and the alternative hypotheses considers that in the case of DM models with QMBE the means are greater. The performed test was the *t-test for paired samples*, which is the adequate test in this situation. If the null hypothesis is rejected, then it can be concluded that it can be accepted that the means are higher in the case of *DM models with QMBE*.

Are easy to understand

H_0 : DM models and DM models with QMBE are equally easy to understand

H_1 : DM models with QMBE are easier to understand

Applying the test mentioned above (values in appendix B), H_0 is rejected in favor of H_1 , with the statistical evidence being very significant (p-value < 0.01). Thus there is statistical evidence that DM models with QMBE are easier to understand.

Are easy to use in practice

H_0 : DM models and DM models with QMBE are equally easy to use in practice

H_1 : DM models with QMBE are easier to use in practice

Applying the test mentioned above (values in appendix B), H_0 is rejected in favor of H_1 , with the statistical evidence being significant (p -value < 0.05). Thus there is statistical evidence that DM models with QMBE are easier to use in practice.

Are oriented to business users

H_0 : DM models and DM models with QMBE are equally oriented to business users

H_1 : DM models with QMBE are more oriented to business users

Applying the test mentioned above (values in appendix B), H_0 is rejected in favor of H_1 , with the statistical evidence being significant (p -value < 0.05). Thus there is statistical evidence that DM models with QMBE are more oriented to business users.

Are oriented to business intelligence activities

H_0 : DM models and DM models with QMBE are equally oriented to business intelligence activities

H_1 : DM models with QMBE are more oriented to business intelligence activities

Applying the test mentioned above (values in appendix B), H_0 is rejected in favor of H_1 , with the statistical evidence being significant (p -value < 0.05). Thus there is statistical evidence that DM models with QMBE are more oriented to business intelligence activities.

Its full potential could be completely explored

H_0 : DM models full potential could be equally explored with QMBE and without QMBE

H_1 : DM models full potential could be better explored with QMBE

Applying the test mentioned above (values in appendix B), H_0 is rejected in favor of H_1 , with the statistical evidence being significant (p-value < 0.05). Thus there is statistical evidence that DM models full potential could be better explored with QMBE.

Help decision making

H_0 : DM models and DM models with QMBE equally help decision making

H_1 : DM models with QMBE better help decision making

Applying the test mentioned above (values in appendix B), H_0 is rejected in favor of H_1 , with the statistical evidence being significant (p-value < 0.05). Thus there is statistical evidence that DM models with QMBE better help decision making.

Bring benefits to Higher Education Institutions

H_0 : DM models and DM models with QMBE equally bring benefits to Higher Education Institutions

H_1 : DM models with QMBE bring more benefits to Higher Education Institutions

Applying the test mentioned above (values in appendix B), H_0 is rejected in favor of H_1 , with the statistical evidence being significant (p-value < 0.05). Thus there is statistical evidence that DM models with QMBE bring more benefits to Higher Education Institutions.

Bring benefits to organizations, in general

H₀: DM models and DM models with QMBE equally bring benefits to organizations, in general

H₁: DM models with QMBE bring more benefits to organizations, in general

Applying the test mentioned above (values in appendix B), H₀ is rejected in favor of H₁, with the statistical evidence being significant (p-value < 0.05). Thus there is statistical evidence that DM models with QMBE bring more benefits to organizations, in general.

From these statistical tests it can be concluded that the means differences are significant (p-value < 0.05) for all cases, except for the first sentence that is very significant (p-value < 0.01). Hence, it can be conclude that, accordingly to the respondents' answers, DM models with QMBE are easier to understand, easier to use in practice, more oriented to business users, more oriented to BI activities, its full potential could be completely explored, it could better help decision making, could bring more benefits to Higher Education Institutions, and could bring more benefits to organizations in general (Table 9).

Table 9 – Comparison of respondents' opinions about using DM vs using DM with QMBE

Statement	Means of 5-point Likert Scale		Statistical test
	DM models	DM models with QMBE	
Are easy to understand.	3,31	3,94	Very Significant
Are easy to use in Practice.	3,25	3,75	Significant
Are oriented to business users.	3,19	3,75	Significant
Are oriented to Business Intelligence Activities.	3,34	3,94	Significant
Its full potential could be completely explored.	3,13	3,69	Significant
Help decision making.	3,75	4,31	Significant
Bring benefits to Higher Education Institutions.	3,56	4,06	Significant
Bring benefits to organizations, in general.	3,75	4,19	Significant

Part 3 - Question 3

Most of the respondents will consider using QMBE in their organizations (question 3 of part 3) as can be concluded from the graph presented in Figure 50. Three of the respondents answered “Certainly”, and seven of the respondents answered “Possibly yes”. These corresponds to 62,5%. Only one of the respondents answered “Not at all”.

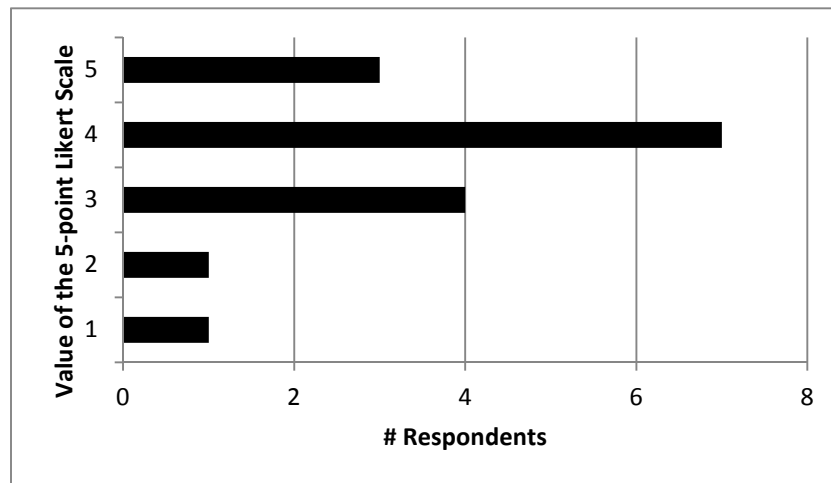


Figure 50 – Respondents’ opinion about adopting QMBE in their organizations (Graph)

Part 3 - Question 4

Some of the respondents presented some comments reinforcing that they consider QMBE and the idea useful and interesting. Other respondents presented comments that were not directly related with the project, but with DM and BI in general.

Part 3 - Question 5

Ten of the respondents (62,5%) declared to be interested in knowing the evaluation results of this study. This can be considered as an additional indicator of the respondents’ interest.

11.3 Some brief considerations

The obtained results are considered promising. The questionnaire respondents expressed a very positive opinion about QMBE. Highlighting question 2 of part 3 of the questionnaire, statistical evidence was given to the fact that the respondents considerer that the use of QMBE provide an enhancement to the use of DM models in the context of BI.

From the analysis of the results obtained with the questionnaire, and from the considerations presented in Section 11.1, it can be concluded that the answer to the second question referred in chapter 7 “Is it an improvement?” is “Yes, it is.”

PART IV – FINAL CONCLUSIONS

In part IV final conclusions are presented. It begins with discussion and related work in chapter 12. In chapter 13 conclusion and future research directions are presented, introducing the thesis' contributions, in Section 13.1, and a critical reflection about the results obtained so far and future work in Section 13.2.

12 Discussion and Related Work

The success of Codd's relational model for databases led to the development of several languages that allow data manipulation and that also allow obtaining quick answers to ad-hoc business questions through queries on the data stored in databases. In relational databases, data is stored in tables that are also called relations. Initially, two formal languages were defined: Relational Algebra and Relational Calculus. Since that time, several languages were developed in order that users could access the data stored in databases. QBE languages were developed with success. Since the first developments (Zloof, 1975; Zloof, 1977), many advances occurred in the area, and the philosophy behind Query-By-Example (QBE) is being applied in several distinct areas (Braga, Campi, Ceri & Spoletini, 2007; Ferreira, Cruz & Henriques, 2009; Malerba, Appice & Vacca, 2002; Papadias & Sellis, 1995). QBE languages are nowadays available in several Relational Database Management Systems (RDBMS). Those languages allow business users to directly manipulate data without the need of developing programming skills. It can be said that a QBE language is business oriented, and it is iterative and interactive in nature, since it allows users obtaining answers to ad-hoc business questions that can be directly converted into QBE queries. Business users frequently pose questions that can be answered through queries to a database. Those queries allow the selection of the database's data that grant the answer to the referred business questions. The use of QBE languages by business users to directly obtain those answers is a usual practice in organizations nowadays. The language presented in this thesis, named Query-Models-by-Example (QMBE) is a Data Mining (DM) language developed as an extension of a QBE language, thus being business oriented, interactive and iterative in nature. This language allows business users to query the data mining models as well as the data, both stored in database tables. As a consequence of being an extension of a QBE language, this new DM language is iterative and interactive in nature. It allows business users to answer ad-hoc business questions through queries on data or/and on DM models. QMBE allows business users to directly access and manipulate DM models. The novelty of the QMBE language is that it is oriented to business users and to BI activities. This kind of approach allows business users to directly access and manipulate data and models. This will bring DM to the front line business users, like other Business Intelligence (BI) tools, allowing them to completely explore DM potential value.

The Inductive Database (IDB) framework is being used by researchers in order to achieve DM standards similar to the ones defined for the relational model for databases. According to the IDB framework, data and DM models are both stored in the database and can be accessed and manipulated at the same level (De Raedt, 2003; Dzeroski, 2007; Imielinski & Mannila, 1996). The research presented in this thesis uses this framework in the context of BI, obtaining the concept of Inductive Data Warehouse (IDW), which is a Data Warehouse (DW) storing both data and DM models. Data and DM models that are stored in the IDW can be manipulated using an inductive DM language.

Several approaches have been proposed for the definition of DM languages. In the literature some language specifications can be found, namely, DMQL (Han, Fu, Wang, Koperski & Zaiane, 1996), MINE RULE (Meo, Psaila & Ceri, 1998), MSQL (Imielinski & Virmani, 1999), SPQL (Bonchi, Giannotti, Lucchesse, Orlando, Perego & Trasarti, 2007), KDDML (Romei, Ruggieri & Turini, 2006), XDM (Meo & Psaila, 2006), RDM (De Raedt, 2002), among others. Despite the importance of the referred languages, they are not business oriented and they are not oriented to the diverse BI activities. The language introduced in this thesis differs from the ones mentioned above in the way that it is oriented to business users and to BI activities.

Research developed in (Wang & Wang, 2008) is aligned with our research taking into account that they consider that business users have a crucial role in the development and analysis of DM models. However, they consider a different approach. They present a model that allows knowledge sharing among business insiders and DM specialists. They argue that this model can make DM more relevant to BI. The research presented in this thesis focus on making DM models to be directly manipulated by business users. It is considered that this can conduct to an understanding of DM models by business users, helping them on the decision making process. This can be done by means of a DM Language that allows business users to query data and models.

The research presented in this thesis is a step in inducting business users of BI systems into DM models and comprises an important contribution towards the goal of binding DM to final business users of BI systems. A long road is yet to be covered but we believe that this research could be an important contribution since it demonstrates that it is possible for business users to directly access, better manipulate and better understand DM models, instead of depending on reports from DM specialists.

13 Conclusion and Future Research Directions

The authors introduced the concept of Inductive Data Warehouse (IDW) and presented a new DM language, Query-Models-By-Example (QMBE), which is iterative, and interactive in nature.

An IDW stores both data and data mining models in database tables. This way, both data and data mining models are stored at the same level. Thus, data mining models can be manipulated at the same level than data, thus allowing business users to manipulate directly data mining models, in the same way that is done with data.

QMBE is a declarative language. This means that the users define “what to do” instead of defining “How to do it”. QMBE is also a high level language, since it is closer to natural languages. These aspects make the language user-friendly, so it is easily used by business users.

Business questions can be converted into queries in the QMBE language, thus it is oriented to BI activities and to BI business users. This will allow business users to directly manipulate DM models, as well as data, thus bringing DM into the final business users, allowing to increase DM potential to attain BI’s high potential business value. This was achieved through the design and implementation of a BI system architecture including DM.

QMBE language is extensible and flexible because new models can easily be made accessible to business users, since it is only necessary to include a new database table for each new model obtained from DM application. The concept is also context independent, since it can also be applied to business situations other than the ones presented in this thesis (a higher education institution). This is due to the fact that the QMBE language is independent from the considered database.

From these considerations, it can be concluded that the research goals were achieved through the introduction of the IDW concept, and the corresponding inductive language, which is QMBE.

13.1 Thesis Contributions

The main contribution of this thesis is to verify the viability of allowing business users to directly manipulate DM models and thus providing the possibility of exploring the potential value of applying DM in the context of BI. This was achieved through the development of two new important concepts, which are themselves two important contributions of this research:

- One of those concepts is the concept of Inductive Data Warehouse (IDW). To put it in a simplified way, an IDW is a DW that contains both data and DM models stored in tables of the DW;
- The other concept is a new DM language: QMBE. This language was developed as an extension of QBE languages, and allows users access to data and DM models, both stored in tables of the DW.

The application of IDW and QMBE provides the possibility of business users to manipulate directly data mining models, thus allowing them to explore the potential value of applying DM in the context of BI.

Another contribution is the design and presentation of an architecture for BI systems that includes the use of DM. This architecture incorporates the new concepts of IDW and of the correspondent inductive DM language, namely QMBE.

Another contribution concerns the use of the Design Science Research (DSR) framework. Despite the growing acceptance of the use of this framework in IS research, there cannot be found many studies using this framework. Therefore, it is considered that the research presented in this thesis can help bringing new insights to research based in the DSR framework.

DSR involves two types of contributions. On one hand, the applications to the environment that led to the satisfaction of the business needs identified in the problem definition. On the other hand the additions to the knowledge base. A summary of both types of contributions that were achieved with this research are presented in Figure 51. The contributions concerning the applications to the environment are:

- the design and implementation of an architecture for a BI system including DM usage;
- allow business users to directly manipulate DM models;
- providing the possibility of business users to explore the potential value of applying DM in the context of BI.

The contributions concerning the additions to the knowledge base are:

- the concept of IDW;
- a new DM language: QMBE.

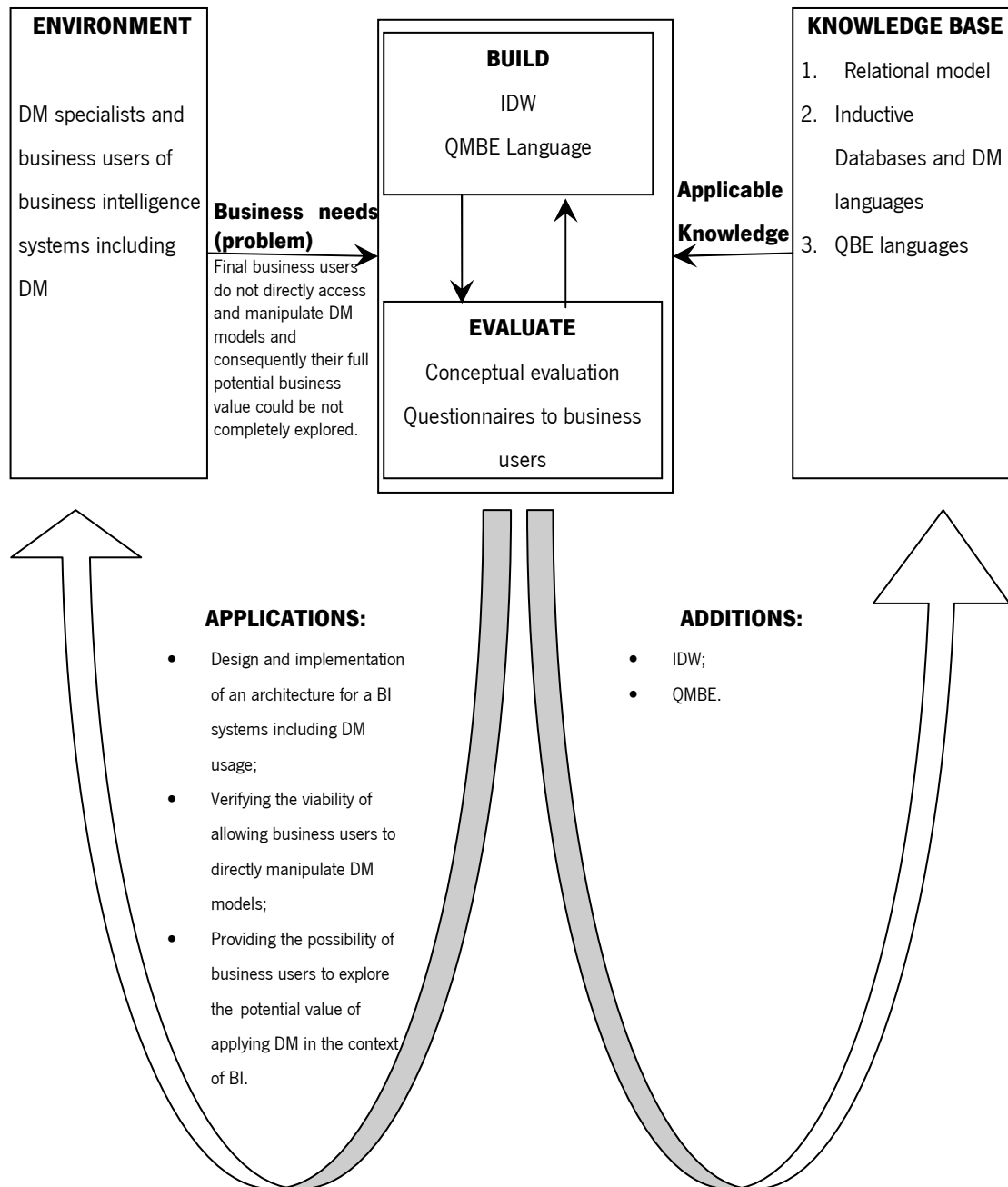


Figure 51 – Research contributions

13.2 Critical reflection about the obtained results and future work

The presented architecture (Figure 22), which includes the use of an IDW and of an inductive DM language, QMBE, was implemented as a prototype. The implementation conducted to promising results. Despite the promising results, several limitations can be pointed out. Following are presented some of those limitations, and future work that can be developed in order to overcome those limitations is pointed out.

One limitation is related with the fact that the system is not completely automated, at the moment. There is the need to develop some additional modules that allow the automatic feed of the DM models stored in the IDW. Existing standards for DM developed by the industry will surely help in this process. The periodicity of this process will be a crucial issue to consider.

The quality of the generated models was not a concern during this research, since this was not considered as an important issue to address at the moment. But it is another crucial issue to address in the future, since only relevant DM models incorporates business value to decision makers, and consequently, only those allow business users to understand the importance of using DM in BI environments to improve decision making.

Another limitation is that only classification rules are addressed at the moment. The application domain focuses mainly on two DM tasks: classification and clustering. Classification can be addressed through classification rules, which can be easily understand by business users. Thus, classification rules were used in this research. In the future, also clustering can be addressed. Like classification, clustering has several important applications in the context of BI systems and is, consequently, important to include it in the system. In this case, it will be challenging to envisage how to store the models in an IDW table. The use of more than one model stored in the same IDW will be considered in this situation.

User interface is also a concern. At the moment the interface only has basic capabilities, and is not very robust. Thus, improvements are planned.

Performance tests will be implemented, in order to better understand the way business users can take advantage of the use of these concepts in concrete situations, allowing to improve their application. The tests will allow business users to perform several tasks that will comprise the definition of business questions and the acquisition of answers to those business questions through the use of the QMBE language, similar to the ones included in this thesis. These performance tests will also help in the definition of the interface improvements, because they will allow to find out the business users difficulties with the use of the system.

Future research directions can thus include tests with more than one model table, the inclusion of clustering models, the automation of the system, and the development of a more robust interface.

It is also intended to progress with the implementation of the developed prototype in different domains. This will surely bring new and important inputs to improve the system.

REFERENCES

- Abbasi, A. & Chen, H. (2008). Cybergate: A Design Framework and System for Text Analysis of Computer-Mediated Communication. *MIS Quarterly*, 32(4), 811-837.
- Adomavicius, G., Bockstedt, J. C., Gupta, A. & Kauffman, R.J. (2008). Making Sense of Technology Trends in the Information Technology Landscape: A Design Science Approach. *MIS Quarterly*, 32(4), 779-809.
- Alavi, M. & Leidner, D.E. (2001). Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. *MIS Quarterly*, 25(1), 107-136.
- Apte, C., Grossman, E., Pednault, E. P. D., Rosen, B. K., Tipu, F. A. & White, B. (1999). Probabilistic Estimation-Based Data Mining for Discovering Insurance Risks. *IEEE Intelligent Systems*, 14(6), 49-58.
- Arnott, D. & Pervan, G. (2008). Eight 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 Weghorn, H. & Abraham, A. P. (Eds.), *Proceedings of the IADIS European Conference on Data Mining 2008*, IADIS MULTI Conference on Computer Science and Information Systems, 182-185. Amsterdam, Holland: IADIS Press.
- Azevedo, A. & Santos, M.F. (2009a). Business Intelligence: State of the Art, Trends, and Open Issues. In Liu, K. (Ed.), *Proceedings of the International Conference on Knowledge Management and Information Sharing (KMIS 2009)*, International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management (IC3K), 296-300. Funchal, Portugal: INSTICC.
- Azevedo, A. & Santos, M.F. (2009b). An Architecture for an Effective Usage of Data Mining in Business Intelligence Systems. In Soliman, K. S. (Ed.), *Knowledge Management and Innovation in Advancing Economies: Analyses & Solutions*, Proceedings of 13th IBIMA Conference, 1319-1325. Marrakech, Morocco: IBIMA.
- Azevedo, A. & Santos, M.F. (2011). A Perspective on Data Mining Integration with Business Intelligence. In Kumar, A. (Ed.), *Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains* (pp.109-129). Hershey, NY: IGI Publishing.
- Benoît, G. (2002). Data Mining. *Annual Review of Information Science and Technology (ARIST)*, 36(1), 265-310.

- Bhandari, I., Colet, E., Parker, J., Pines, Z., Pratap, R. & Ramanujam, K. (1997). Advanced Scout: Data Mining and Knowledge Discovery in NBA Data. *Data Mining and Knowledge Discovery*, 1(1), 121-125.
- Bollacker, K. D., Lawrence, S. & Giles, C.L. (2000). Discovering Relevant Scientific Literature on the Web. *IEEE Intelligent Systems*, 15(2), 42-47.
- Bonchi, F.; Giannotti, F.; Lucchese, C.; Orlando, S.; Perego, R. & Trasarti, R. (2007). On Interactive Pattern Mining from Relational Databases. In Dzeroski, S. & Struyf, J. (Eds.), *Lecture Notes on Computer Science: Vol. 4747. Knowledge Discovery in Inductive Databases - 5th International Workshop, KDID 2006* (pp. 42-62). Berlin, Heidelberg: Springer-Verlag.
- Botta, M.; Boulicaut, J.; Masson, C. & Meo, R. (2004). Query Languages Supporting Descriptive Rule Mining: A comparative Study. In Meo, R. ; Lanzi, P. L. & Klemettinen, M. (Eds.), *Lecture Notes on Artificial Intelligence: Vol. 2682. Database Support for Data Mining Applications - Discovering Knowledge with Inductive Queries* (pp. 24-51). Berlin, Heidelberg: Springer-Verlag.
- Boulicaut, J.; Klemettinen, M. & Mannila, H. (1999). Modeling KDD Processes Within the Inductive Database Framework. In Mohania, M. & Tjoa, A. M. (Eds.), *Lecture Notes on Computer Science: Vol. 1676. Data Warehousing and knowledge Discovery - 1st International Conferense DaWak99* (pp. 193-202). Berlin, Heidelberg: Springer-Verlag.
- Brachman, R. J. & Anand, T. (1996). The Process of Knowledge Discovery in databases. In Fayyad, U. M. , Piatetski-Shapiro, G. , Smyth, P. & Uthurusamy, R. (Eds.), *Advances in knowledge discovery and data mining* (pp.37-57). Menlo Park, CA: AAAI Press/The MIT Press.
- Braga, D., Campi, A., Ceri, S. & Spoletini, P. (2007). XQuery Layers. *SIGMOD Record*, 36(1), 25-30.
- Brobst, S. & Pareek, A. (2009). New Trends in Data Acquisition Services for the Real-Time Enterprise. *Business Intelligence Journal*, 14(1), 52-58.
- Calders, T., Lakshmanan, L. V. S., Ng, R. T. & Paredaens, J. (2006). Expressive Power of an Algebra for Data Mining. *ACM Transactions on Database Systems*, 31(4), 1169-1214.
- 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.
- Carlsson, S. A. (2006). Towards an Information Systems Design Research Framework: A Critical Realist Perspective. *Proceedings of the First International Conference on Design Science in Information Systems and Technology - DESRIT 2006*, 192-212.
- Catania, B.; Maddalena, A.; Mazza, M.; Bertino, E. & Rizzi, S. (2004). A Framework for Data Mining Pattern Management. In Boulicaut, J. ; Esposito, F. & Giannotti, F. (Eds.), *Lecture Notes on Artificial Intelligence: Vol. 3202. Knowledge Discovery in Databases - 8th European Conference on Principles and Practice of Knowledge Discovery in Databases - PKDD2004* (pp. 87-98). Berlin, Heidelberg: Springer-Verlag.
- Chan, P. K., Fan, W., Prodromidis, A. L. & Stolfo, S.J. (1999). Distributed Data Mining in Credit Card Fraud Detection. *IEEE Intelligent Systems*, 14(6), 67-74.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). CRISP-DM 1.0 - Step-by-step data mining guide. Technical Report: 1.0, CRISP-DM consortium.
- Chen, M., Han, J. & Yu, P.S. (1996). Data Mining: An Overview from a Database Perspective. *IEEE transactions on Knowledge and Data Engineering*, 8(6), 866-883.
- Cheng, H., Lu, Y. & Sheu, C. (2009). An Ontology-Based Business Intelligence Application In A Financial Knowledge Management System. *Expert Systems with Applications*, 36(2), 3614-3622.
- Chiang, I., Shieh, M., Hsu, J. Y. & Wong, J. (2005). Building a Medical Decision Support System for Colon Polyp Screening by Using Fuzzy Classification Trees. *Applied Intelligence*, 22(1), 61-75.
- 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.
- Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM*, 13(6), 377-387.
- Codd, E. F. (1971). A Data Base Sublanguage Founded on the Relational Calculus. *Proceedings of the 1971 ACM SIGFIDET (SIGFIDET '71) - Workshop on Data Description, Access, and Control*, 35-68.

- Codd, E. F. (1982). Relational Database: a Practical Foundation for Productivity. *Communications of the ACM*, 25(2), 109-117.
- Corchado, J. M., Mata, A., Paz, F. D. & Pozo, D.D. (2008). A Case-Based Reasoning System to Forecast the Presence of Oil Slicks. In Weghorn, H. & Abraham, A. P. (Eds.), *Proceedings of the IADIS European Conference on Data Mining 2008*, IADIS Multi Conference on Computer Science and Information Systems, 3-10. Amsterdam, Holland: IADIS Press.
- Cox, K. C., Eick, S. G., Wills, G. J. & Brachman, R.J. (1997). Visual Data Mining: Recognizing Telephone Calling Fraud. *Data Mining and Knowledge Discovery*, 1(2), 225-231.
- Data Mining Group (2009). Predictive Model Markup Language (PMML). Data Mining Group Portal, 2009. Retrieved August, 1st 2009, from <http://www.dmg.org/>.
- Date, C. J. (2004). *An Introduction to Database Systems*. Upper Saddle River, New Jersey: Pearson Education.
- Davenport, T. H. (2010). Business Intelligence and Organizational Decisions. *International Journal of Business Intelligence Research*, 1(1), 1-12.
- 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.
- Dzienciowski, K. & Kina, D. (2008). Data Mining in Marketing Acquisition Campaigns. In Weghorn, H. & Abraham, A. P. (Eds.), *Proceedings of the IADIS European Conference on Data Mining 2008*, IADIS Multi Conference on Computer Science and Information Systems, 173-175. Amsterdam, Holland: IADIS Press.
- Eckerson, W. W. (2008). Q&A: Pervasive Business Intelligence. *Business Intelligence Journal*, 13(3), 48-50.
- Eckerson, W. W. (2009). Performance Management Strategies. *Business Intelligence Journal*, 14(1), 24-27.

- Elbashir, M. Z., Collier, P. A. & Davern, M.J. (2008). Measuring the effects of business intelligence systems: The relationship between business process and organizational performance. *International Journal of Accounting Information Systems*, 9(3), 135-153.
- Elmasri, R. & Navathe, S.B. (2007). *Fundamentals of Database Systems*. Upper Sadle River, New Jersey: Pearson Education.
- Ezawa, K. J. & Norton, S.W. (1996). Constructing Bayesian Networks to Predict Uncollectible Telecommunications Accounts. *IEEE Expert*, 11(5), 45-51.
- Fawcett, T. & Provost, F. (1997). Adaptive Fraud Detection. *Data Mining and Knowledge Discovery*, 1(3), 291-316.
- Fayyad, U. M. (1996). Data Mining and Knowledge Discovery: Making Sense Out of Data. *IEEE Expert*, 11(5), 20-25.
- Fayyad, U. M., Djorgovski, S. G. & Weir, N. (1996). Automating the analysis and Cataloging of Sky Surveys. In Fayyad, U. M. , Piatetski-Shapiro, G. , Smyth, P. & Uthurusamy, R. (Eds.), *Advances in knowledge discovery and data mining* (pp.471-493). Menlo Park, California: AAAI Press/The MIT Press.
- 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.
- Ferreira, F. X., Cruz, D. & Henriques, P.R. (2009). A Query by Example Approach for XML Querying. In Rocha, Á. , Restivo, Francisco, Reis, L. P. & Torrão, S. (Eds.), *WSIAI'09 - Workshop em Sistemas Inteligentes e Aplicações, CISTI'09 - 4ª Conferência Ibérica em Sistemas e Tecnologias da Informação*, 611-614. Póvoa do Varzim, Portugal: AISTI.
- 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.
- Fung, G. & Stoeckel, J. (2007). SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information. *Knowledge and Information Systems*, 11(2), 243-258.
- Gago, P. & Santos, M.F. (2008). Towards an Intelligent Decision Support System for Intensive Care Units. In Okun, O. & Valentini, G. (Eds.), *Workshop on Supervised and Unsupervised*

- Ensemble Methods and their Applications*, Proceedings of the 18th European Conference on Artificial Intelligence, 21-25. Patras, Greece: WSEAS.
- Gago, P., Fernandes, C., Pinto, F. & Santos, M.F. (2009). INTCare: On-line Knowledge Discovery in the Intensive Care Unit. *INES'09 Proceedings of the IEEE 13th international conference on Intelligent Engineering Systems*, 143-148.
- Gao, J., Cao, Y., Qi, Y. & Hu, J. (2005). Building Innovative Representations of DNA Sequences to Facilitate Gene Finding. *IEEE Intelligent Systems*, 20(6), 34-39.
- Gerber, L. & Fernandes, A.A.A. (2004). An Abstract Algebra for Knowledge Discovery in Databases. In Benczúr, A. ; Demetrovics, J. & Gottlob, G. (Eds.), *Lecturer Notes in Computer Science: Vol. 3255. Advances in Database and Information Systems* (pp. 83-98). Berlin, Heidelberg: Springer-Verlag.
- Ghosh, J. & Strehl, A. (2005). Clustering and Visualization of Retail Market Baskets. In Pal, N. R. & Jain, L. (Eds.), *Advanced Techniques in Data Mining and Knowledge Discovery* (pp.75-102). London, UK: Springer-Verlag.
- Gokhale, M. & Aslandogan, Y.A. (2003). A Visualization Oriented Data Mining Tool for Biomedical Images. In Smari, W. W. & Memon, A. M. (Eds.), *IEEE International Conference on Information Reuse and Integration, 2003, IRI 2003*, 219-226. Las Vegas, NV: IEEE Press.
- Golfarelli, M., Rizzi, S. & Cella, I. (2004). Beyond Data Warehousing: What `s Next in Business Intelligence. *DOLAP '04 Proceedings of the 7th ACM international workshop on Data warehousing and OLAP*, 1-6.
- Gurbaxani, B. & Mallick, P. (2005). Fiding Protein Domain Boundaries: an Automated, Non-Homology-Based Method. *IEEE Intelligent System*, 20(6), 26-33.
- Han, J. & Kamber, M. (2006). *Data Mining: concepts and Techniques*. San Francisco, CA: Morgan Kaufman Publishers.
- Han, J., Fu, Y., Wang, W., Koperski, K. & Zaiane, O. (1996). DMQL: A Data Mining Query Language for Relational Databases. *Proceedings of the SIGMOD'96 Workshop on Research Issues on Data Minining and Knowledge Discovery (DMKD'96)*, 27-34.
- Hand, D., Mannila, H. & Smyth, P. (2001). *Principles of Data Mining*. Cambridge, Massachusetts: The MIT Press.
- 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.

- Herschel, R. T. & Jones, N.E. (2005). Knowledge Management and Business Intelligence: the Importance of Integration. *Journal of Knowledge Management*, 9(4), 45-55.
- Hevner, A. & Chatterje, S. (2010). Design Research in Information Systems: Theory and Practice. In Sharda, R. & Vob, S. (Eds.), *Integrated Series in Information Systems: Vol. 22*. (pp. 1-320). Berlin, Heidelberg: Springer-Verlag.
- Hevner, A. R., March, S. T., Park, J. & Ram, S. (2004). Design Science Research in Information Systems Research. *MIS Quarterly*, 28(1), 75-105.
- Hirschheim, R. A. (1985). Information Systems Epistemology: an Historical Perspective. In Munford, E. , Hirschheim, R. , Fitzgerald, G. & Wood-Harper, T. (Eds.), *Research Methods in Information Systems* (pp.13-36). Amsterdam, North Holland: Elsevier Science publishing.
- Hobek, R., Ariyachandra, T. & Frolick, M.N. (2009). The Importance of Soft Skills in Business Intelligence Implementations. *Business Intelligence Journal*, 14(1), 28-36.
- Hoffman, T. (2009). 9 Hottest Skills for '09. *Computer World*, January 1(1), 26-27.
- Hsu, C., Chung, H. & Huang, H. (2004). Mining Skewed and Sparse Transaction Data for Personalized Shopping Recommendation. *Machine Learning*, 57(1), 35-59.
- Hu, X. (2005). A Data Mining Approach for Retailing Bank Customer Attrition Analysis. *Applied Intelligence*, 22(1), 47-60.
- Imielinski, T. & Mannila, H. (1996). A Database Perspective on Knowledge Discovery. *Communications of the ACM*, 39(11), 58-64.
- Imielinski, T. & Virmani, A. (1999). MSOL: 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.
- John, G. H., Miller, P. & Kerber, R. (1996). Stock Selection Using Rule Induction. *IEEE Expert*, 11(5), 52-58.
- KDNuggets (2011). Software Suites for Data Mining, Analytics, and Knowledge Discovery. Data Mining Community's Resource, 1. Retrieved April 2011, from <http://www.kdnuggets.com/software/suites.html>.
- Kimball, R. & Ross, M. (2002). *The Data Warehouse Toolkit - The Complete Guide to Dimensional Modeling*. Hoboken, NJ: John Wiley and Sons.

- Klawans, B. (2008). Embedded or Conventional BI: Determining the Right Combination of BI for Your Business. *Business Intelligence Journal*, 13(1), 30-36.
- 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, Heidelberg: Berlin-Verlag.
- Kriegel, H., Borgwardt, K. M., Kröger, P., Pryakhin, A., Schubert, M. & Zimek, A. (2007). Future Trends in Data Mining. *Data Mining and Knowledge Discovery*, 15(1), 87-97.
- Kudyba, S. & Hoptroff, R. (2001). *Data Mining and Business Intelligence: a Guide to Productivity*. Hershey, NY: IGI Publishing.
- Kumar, A. V. S. (2011). *Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains*. Hershey, New York: IGI Publishing.
- König, A. & Gratz, A. (2005). Advanced Methods for the Analysis of Semiconductor Manufacturing Process Data. In Pal, N. R. & Jain, L. (Eds.), *Advanced Techniques in Data Mining and Knowledge Discovery* (pp.27-74). London, UK: Springer-Verlag.
- Lappas, G. (2009). Machine Learning and Web Learning: Methods and Applications in Societal Benefit areas. In Rahman, H. (Ed.), *Data Mining Applications for Empowering Knowledge Societies* (pp.76-95). Hershey, New York: IGI Publishing.
- Larose, D. T. (2005). *Discovering Knowledge in Data: an Introduction to Data Mining*. Hoboken, New Jersey: John Wiley & Sons.
- Lee, J., Wyner, G. M. & Pentland, B.T. (2008). Process Grammar as a Tool for Business Process Design. *MIS Quarterly*, 32(4), 757-778.
- Li, S., Shue, L. & Lee, S. (2008). Business Intelligence Approach to Supporting Strategy-making of ISP Service Management. *Expert Systems with Applications*, 35(3), 739-754.
- Liabotis, I., Theodoulidis, B. & Saraaee, M. (2006). Improving Similarity Search in Time Series Using Wavelets. *International Journal of Data Warehousing and Mining*, 2(2), 55-81.
- Liebowitz, J. (2006). *Strategic Intelligence: Business Intelligence, Competitive Intelligence, and Knowledge Management*. Boca Raton, FL: Auerbach Publications.
- Lin, C., Pu, H. & Lee, Y. (2005). Satellite Image Classification Using Cascaded Architecture of Neural Fuzzy Network. In Pal, N. R. & Jain, L. (Eds.), *Advanced Techniques in Data Mining and Knowledge Discovery* (pp.211-231). London, UK: Springer-Verlag.

- Lin, Y., Tsai, K., Shiang, W., kuo, T. & Tsai, C. (2009). Research on using ANP to establish a performance assessment model for business intelligence systems. *Expert Systems with Applications*, 36(2), 4135-4146.
- Linoff, G. S. (2008). Survival Data Mining Using Relational Databases. *Business Intelligence Journal*, 13(3), 20-30.
- Luck, D. (2009). The Importance of Data Within Contemporary CRM. In Rahman, H. (Ed.), *Data Mining Applications for Empowering Knowledge Societies* (pp.96-109). Hershey, New York: IGI Publishing.
- Lunger, K. (2008). Debunking Three Myths of Pervasive Business Intelligence: How to Create a Truly Democratic BI Environment. *Business Intelligence Journal*, 13(4), 38-41.
- Lunh, H. P. (1958). A Business Intelligence System. *IBM Journal of Research and Development*, 2(4), 314-319.
- Létourneau, S., Famimi, F. & Matwin, S. (1999). Data Mining to Predict Aircraft Component Replacement. *IEEE Intelligent Systems*, 14(6), 59-65.
- Malerba, D., Appice, A. & Vacca, N. (2002). SDMOQL: An OQL-based Data Mining Query Language for Map Interpretation Tasks. *Proceedings of the EDBT 2002 Workshop on Database Technologies for Data Mining*, 3-18.
- Mannila, H. (2000). Theoretical Frameworks for Data Mining. *SIGKDD Explorations*, 1(2), 30-32.
- March, S. T. & Hevner, A.R. (2007). Integrated decision support systems: A data warehousing perspective. *Decision Support Systems*, 43(3), 1031-1043.
- March, S. T. & Smith, G.F. (1995). Design Science and Natural Science Research on Information Technology. *Decision Support Systems*, 15(4), 251-266.
- March, S. T. & Storey, V.C. (2008). Design Science in the Information Systems Discipline: An Introduction to the Special Issue on Design Science Research. *MIS Quarterly*, 32(4), 725-730.
- Markov, Z. & Larose, D.T. (2007). *Data mining the Web: uncovering patterns in Web content, structure, and usage*. Hoboken, New Jersey: Wiley-Interscience.
- Martin, J., Gibrat, J. & Rodolphe, F. (2005). Choosing the Optimal Hidden Markov Model for Secondary-Structure Prediction. *IEEE Intelligent Systems*, 20(6), 19-25.

- McKay, J. & Marshall, P. (2005). A Review of Design Science in Information Systems. *Proceedings of the 16th Australasian Conference on Information Systems - ACIS 2005*, 1-11.
- McKnight, W. (2002). Bringing 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/20021101/5980-1.html>.
- McKnight, W. (2003). Bringing Data Mining to the Front Line, Part 2. *Information Management magazine*, November(2002), Retrieved on July, 16th 2009 at <http://www.information-management.com/issues/20021101/5980-1.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.
- Michalewicz, Z., Schmidt, M., Michalewicz, M. & Chiriac, C. (2007). *Adaptive Business Intelligence*. Heidelberg, Berlin: Springer-Verlag.
- Mielikäinen, T. (2004). Inductive Databases as Ranking. In Kambayashi, Y. ; Mohania, M. & Wöb, W. (Eds.), *Lecture Notes on Computer Science: Vol. 3181. Data Warehousing and knowledge Discovery - 6th International conference DaWak2004* (pp. 149-158). Berlin, Heidelberg: Springer-Verlag.
- Moss, L. T. & Shaku, A. (2003). *Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications*. Upper Saddle River, NJ: Pearson Education.
- Myatt, G. J. (2007). *Making Sense of Data - A Practical Guide to Exploratory Data Analysis and Data Mining*. Hoboken, New Jersey: John Wiley & Sons.
- Negash, S. (2004). Business Intelligence. *Communications of the Association for Information Systems*, 13(1), 177-195.
- Nemati, H. R., Steiger, D. M., Iyer, L. S. & Herschel, R.T. (2002). Knowledge Warehouse: an Architectural Integration of Knowledge Management, Decision Support, Artificial Intelligence and Data Warehousing. *Decision Support Systems*, 33(2), 143-161.

- Nijssen, S. & De Raedt, L. (2007). IQL: A Proposal for an Inductive Query Language. In Dzeroski, S. & Struyf, J. (Eds.), *Lecture Notes in Computer Science: Vol. 4747. Knowledge Discovery in Inductive Databases - 5th International Workshop, KDID 2006* (pp. 189-209). Berlin, Heidelberg: Springer-Verlag.
- Nlenanya, I. (2009). Building an Environmental GIS Knowledge Infrastructure. In Rahman, H. (Ed.), *Data Mining Applications for Empowering Knowledge Societies* (pp.262-279). Hershey, New York: IGI Publishing.
- Object Management Group (2008). Knowledge Discovery Model (KDM). KDM Portal, 2006-2008. Retrieved August, 1st 2009, from <http://kdmanalytics.com/kdm/index.php>.
- Pan, J., Yang, Q., Yang, Y., Li, L., Li, F. T. & Li, G.W. (2007). Cost-Sensitive-Data Preprocessing for Mining Customer Relationship Management Databases. *IEEE Intelligent Systems*, 22(1), 46-51.
- Papadias, D. & Sellis, T. (1995). A Pictorial Query-by-Example Language. *Journal of Visual Languages and Computing*, 6(1), 53-72.
- Parsons, J. & Wand, Y. (2008). Using Cognitive Principles to Guide Classification in Information Systems Modeling. *MIS Quarterly*, 32(4), 839-868.
- Pereira, R. H., Azevedo, A. & Castilho, O. (2007). Secretaria On-Line From Iscap: A Case of Innovation. In Nunes, M. B. , Isaías, P. & Barroso, J. (Eds.), *Proceedings of the IADIS International Conference , WWW/Internet 2007*, 301-305. Vila Real, Portugal: IADIS Press.
- Pervan, G. & Arnott, D. (2006). Research in Data Warehousing and Business Intelligence: 1990-2004. In Frédéric, A. ; Brézillon, P. ; Carlsson, S. & Humphreys, P. (Eds.), *Papers from the IFIP WG8.3 International Conference on Creativity and Innovation in Decision Making and Decision Support*,: Vol. 2. Creativity and Innovation in Decision Making and Decision Support (pp. 985-1003). London, UK: Ludic Publishing Ltd.
- Piatetsky-Shapiro, G. (2007). Data Mining and Knowledge Discovery 1996 to 2005: Overcoming the Hype and Moving from "university" to "business" and "analytics". *Data Mining and Knowledge Discovery*, 15(1), 99-105.
- Pinto, F., Gago, P. & Santos, M.F. (2006). Data Mining as a New Paradigm for Business Intelligence in Database Marketing Projects. In Manolopoulos, Y. , Filipe, J. , Constantopoulos, P. & Cordeiro, J. (Eds.), *ICEIS 2006 - Proceedings of the Eighth International Conference on Enterprise Information Systems, Databases and Information Systems Integration*, 144-149. Paphos, Cyprus: INSTICC.

- Pinto, F., Santos, M. F. & Marques, A. (2009). Database Marketing Intelligence Supported by Ontologies. *WSEAS Transactions on Business and Economics*, 6(3), 135-146.
- Power, D. J. (2007). A Brief History of Decision Support System. DSSResources.COM, Version 4.0. Retrieved March 10, from <http://dssresources.com/history/dsshhistory.html>.
- Prado, H. A. & Ferneda, E. (2008). Emerging Technologies of Text Mining: Techniques and Applications. Retrieved , from <http://www.igi-global.com/reference/details.asp?id=6993>.
- Pries-Heje, J. & Baskerville, R. (2008). The Design Theory Nexus. *MIS Quarterly*, 32(4), 731-755.
- Quintela, H.; Santos, M. F. & Cortez, P. (2007). Real-Time Intelligent Decision Support System for Bridges Structures Behavior Prediction. In Neves, J. ; Santos, M. F. & Machado, J. (Eds.), LNAI: Vol. 4874. Proceedings of the 13th Portuguese Conference on Artificial Intelligence, EPIA 2007 (pp. 124-132). Berlin Heidelberg, Germany: Springer-Verlag.
- Rahman, H. (2009). Prospects and Scopes of Data Mining Applications in Society Development Activities. In Rahman, H. (Ed.), *Data Mining Applications for Empowering Knowledge Societies* (pp.162-188). Hershey, New York: IGI Publishing.
- Raisinghani, M. (2004). *Business Intelligence in the Digital Economy: Opportunities, Limitations and Risks*. Hershey, NY: IGI Publishing.
- Rantza, R. (2004). Frequent Itemset Discovery with SQL Using Universal Quantification. 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. 194-213). Berlin, Heidelberg: Springer-Verlag.
- Richardson, J., Schlegel, K. & Hostmann, B. (2009). Magic Quadrant for Business Intelligence Platforms - 2009. Core Research Note: G00163529, Gartner.
- Richardson, J., Schlegel, K., Hostmann, B. & McMurchy, N. (2008). Magic Quadrant for Business Intelligence Platforms - 2008. Core Research Note: G00154227, 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.
- Sallam, R., Hostman, B., Richardson, J. & Bitterer, A. (2010). Magic Quadrant for Business Intelligence Platforms 2010. Core Research Note: G00173700, Gartner.
- Salzberg, S. L. (1999). Gene Discovery in DNA Sequences. *IEEE Intelligent Systems*, 14(6), 44-48.

- Santos, M. F. & Azevedo, C.S. (2005). *Data Mining - Descoberta de Conhecimento em Bases de Dados*. Lisboa, Portugal: FCA - Editora de Informática.
- Santos, M. F.; Cortez, P.; Pereira, J. & Quintela, H. (2006). Corporate Bankruptcy Prediction Using Data Mining Techniques. In Zanasi, A. ; Brebbia, C. A. & Ebecken, N. F. F. (Eds.), *WIT Transactions on Information and Communication Technologies: Vol. 37. Data Mining VII: Data, Text and Web Mining and their Business Applications* (pp. 349-357). Southampton, UK: WIT Press.
- Santos, M. F.; Cortez, P.; Quintela, H. & Pinto, F. (2005). A Clustering Approach for Knowledge Discovery in Database Marketing. In Zanasi, A. ; Brebbia, C. A. & Ebecken, N. F. F. (Eds.), *WIT Transactions on Information and Communication Technologies: Vol. 35. Data Mining VI: Data, Text and Web Mining and their Business Applications* (pp. 367-376). Southampton, UK: WIT Press.
- Santos, M. F.; Cortez, P.; Quintela, H.; Neves, J.; Vicente, H. & Arteiro, J. (2005). Ecological Mining - A Case Study on Dam Water Quality. In Zanasi, A. ; Brebbia, C. A. & Ebecken, N. F. F. (Eds.), *WIT Transactions on Information and Communication Technologies: Vol. 35. Data Mining VI: Data mining, Text Mining and their Business Applications* (pp. 481-489). Southampton, UK: WIT Press.
- Santos, M., Pereira, J. & Silva, Á. (2005). A Cluster Framework for Data Mining Models: an applications to intensive medicine. In Chen, C. , Filipe, J. , Seruca, I. & Cordeiro, J. (Eds.), *Proceedings of the 7th International Conference on Enterprise Information Systems, ICEIS 2005*, 163-168. Miami, USA: INSTICC.
- 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.
- Scime, A., Murray, G. R., Huang, W. & Brownstein-Evans, C. (2008). Data Mining in the Social Sciences and Iterative Attribute Elimination. In Taniar, D. (Ed.), *Data Mining and Knowledge Discovery Technologies* (pp.308-332). Hershey, New York: IGI Publishing.
- Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R. & Carlsson, C. (2002). Past, Present, and Future of Decision Support Technology. *Decision Support Systems*, 32(2), 111-126.

- Silva, M. S., Câmara, G. & Escada, M.I. (2009). Image Mining: Detecting Deforestation Patterns Through Satellites. In Rahman, H. (Ed.), *Data Mining Applications for Empowering Knowledge Societies* (pp.55-75). Hershey, New York: IGI Publishing.
- Simon, H. A. (1981). *The Sciences of the Artificial*. Cambridge, Massachusetts: Massachusetts Institute of Technology.
- Simoudis, E. (1996). Reality Check for Data Mining. *IEEE Expert*, 11(5), 26-33.
- Smyth, P., Fayyad, U. M., Burl, M. C. & Perona, P. (1996). Modeling Subjective Uncertainty in Image Annotation. In Fayyad, U. M. , Piatetski-Shapiro, G. , Smyth, P. & Uthurusamy, R. (Eds.), *Advances in knowledge discovery and data mining* (pp.519-539). Menlo Park, California: AAAI Press/The MIT Press.
- Steiger, D. M. (2010). Decision Support as Knowledge Creation: a Business Intelligence Design Theory. *International Journal of Business Intelligence Research*, 1(1), 29-47.
- Strenger, L. (2008). Coping with "Big Data" Growing Pains. *Business Intelligence Journal*, 13(4), 45-52.
- Sweets, D. L., Pathak, Y. & Weng, J.J. (1998). An Image Database System with Support for Traditional Alphanumeric Queries and Content-Based Queries by Example. *Multimedia Tools and Applications*, 7(3), 181-212.
- Tadesse, T., Wardlow, B. & Hayes, M.J. (2009). The Application of Data Mining for Drought Monitoring and Prediction. In Rahman, H. (Ed.), *Data Mining Applications for Empowering Knowledge Societies* (pp.280-291). Hershey, New York: IGI Publishing.
- Tang, Z. & MacLennan, J. (2005). *Data Mining with SQL Server 2005*. Indianapolis, IN: Wiley Publishing.
- Thierauf, R. J. (2001). *Effective Business Intelligence Systems*. Westport, CT : Quorum Books.
- Turban, E., Aroson, J. E., Liang, T. & Sharda, R. (2007). *Decision Support and Business Intelligence Systems*. Upper Sadle River, NJ: Pearson Prentice Hall.
- Turban, E., Sharda, R., Aroson, J. E. & King, D. (2008). *Business Intelligence: A Managerial Approach*. Upper Sadle River, NJ: Pearson Prentice Hall.
- Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making*. West Sussex, UK: John Wiley & Sons.
- Wang, H. & Wang, S. (2008). A Knowledge Management Approach to Data Mining Process for Business Intelligence. *Industrial Management & Data Systems*, 108(5), 622-634.

- Watson, H. J. (2009). Bridging the IT/Business Culture Chasm. *Business Intelligence Journal*, 14(1), 4-7.
- Witten, I. H. & Frank, E. (2005). *Data Mining - Practical Machine Learning Tools and Techniques*. San Francisco, CA: Morgan Kaufmann Publishers.
- Wormus, T. (2008). Complex Event Processing: Analytics and Complex Event Processing: Adding Intelligence to the Event Chain. *Business Intelligence Journal*, 13(4), 53-58.
- Wu, C., Yu, L. & Jang, F. (2005). Using Semantic Dependencies to Mine Depressive Symptoms from Consultation Records. *IEEE Intelligent Systems*, 20(6), 50-59.
- Ye, N. (2003). *The Handbook of Data Mining*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Yermish, I., Miori, V., Yi, J., Malhotra, R. & Klimberg, R. (2010). Business Plus Intelligence Plus Technology Equals Business Intelligence. *International Journal of Business Intelligence Research*, 1(1), 48-63.
- Zeller, J. (2007). Business Intelligence: The Chicken or the Egg. *BI Review Magazine*, May 8, 2007. Retrieved February 15, 2009, from <http://www.information-management.com/bissues/20070601/2600340-1.html>.
- Zeller, J. (2008). Business Intelligence: the Road Trip. *Information Management Special Reports*, December 2, 2008. Retrieved February 15, 2009, from http://www.information-management.com/specialreports/2008_112/10002266-1.html?pg=1.
- Zhang, X., Hu, X., Xia, J., Zhou, X. & Achananuparp, P. (2008). A Graph-Based Biomedical Literature Clustering Approach Utilizing Term's Global and Local Importance Information. *International Journal of Data Warehousing and Mining*, 4(4), 84-101.
- Zloof, M. M. (1975). Query-by-Example: the Invocation and Definition of Tables and Forms. *Proceedings of the 1st International Conference on Very Large Databases*, 1-24.
- Zloof, M. M. (1977). Query-by-Example: a data base language. *IBM Systems Journal*, 16(4), 324-343.
- Zloof, M. M. & de Jong, S.P. (1977). The System for Business Automation (SBA): Programming Language. *Communications of the ACM*, 20(6), 385-396.

APPENDIX A – QUESTIONNAIRE

QMBE - Query Models-By-Example

This questionnaire is part of a PhD project. In the research a new language, named Query Models-By-Example (QMBE), was developed as an extension of Query-By-Example (QBE) languages presented in some relational database management systems. The goal is to evaluate the new language (QMBE).

The questionnaire consists of three parts:



- Part 1 - General questions
- Part 2 - QMBE Overview
- Part 3 - Evaluation of the proposed language



Part 1 - General questions

General questions *

1.1: How long have you been using a Business Intelligence System to support decision making in higher education institutions, or other organizations? Please write your answer here:

*** 1.2: How long have you been using Data Mining to support decision making in higher education institutions, or other organizations? Please write your answer here:**

*** 1.3: In your opinion, how important is the use of Data Mining to support decision making? Please choose *only one* of the following:**

- Very important
- Important
- Somewhat important
- Not important
- Not at all important

*** 1.4: Please, explain briefly how do you use Data Mining in higher education institutions, or other organizations, to support decision making.**

(Are there reports from data mining specialists? Do you have direct access to data mining models?..)

Please write your answer here:

Part 2 - Language overview

A film containing a tutorial of the QMBE language is presented.

Please, click **next>>** after watching the movie, so that you can answer to the last group of questions.

2.1:

[Please click here to watch the movie.](#)

Part 3 - Utility of the proposed language

Evaluating the utility of the proposed language

* 3.1: What is your general reaction to the proposed language (QMBE)? Please write your answer here:

⏪
⏩

⏪
⏩

* 3.2: For each of the following sentences express your opinion considering:

- 1 – Strongly disagree
- 2 – Somewhat disagree
- 3 – Neutral/no opinion
- 4 – Somewhat agree
- 5 – Strongly agree

Please choose the appropriate response for each item:

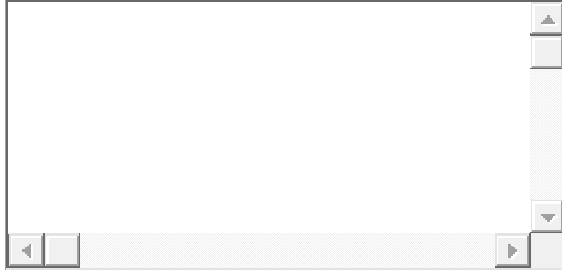
	DM models					DM models with QMBE				
	1	2	3	4	5	1	2	3	4	5
Are easy to understand.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Are easy to use in Practice.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Are oriented to business users.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Are oriented to Business Intelligence Activities.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Its full potential could be completely explored.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Help decision making.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Bring benefits to Higher Education Institutions.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Bring benefits to organizations, in general.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

* 3.3.: Will you consider using the proposed language in your organization to support decision making? Please choose *only one* of the following:

- Certainly
- Possibly yes
- Do not know
- Possibly no
- No at all

Data mining languages for business intelligence

3.4: Please, provide additional suggestions/comments. Please write your answer here:



*** 3.5: Are you interested in the evaluation results?** Please choose *only one* of the following:

- Yes
- No

[Only answer this question if you answered 'Yes' to question '3.5 ']

3.5.1: Please enter your email address. Please write your answer here:

Submit Your Survey.
Thank you for completing this survey..

APPENDIX B – STATISTICAL TESTS

$$H_0: \mu_1 - \mu_2 = 0 (\mu_1 = \mu_2)$$

$$H_1: \mu_1 - \mu_2 > 0 (\mu_1 > \mu_2)$$

H_0 : There is no difference between means μ_1 and μ_2

H_1 : μ_1 is greater than μ_2

Performed test: t-test for paired samples

Question: Are easy to understand.

	<i>Variable 1</i>	<i>Variable 2</i>
Mean	3,9375	3,3125
Variance	0,729166667	0,7625
Observations	16	16
df	15	
Stat t	2,611164839	
p-value	0,009828517	
t value	1,753050325	

Stat t > t value thus Reject H_0 Accept H_1

Question: Are easy to use in Practice.

	<i>Variable 1</i>	<i>Variable 2</i>
Mean	3,75	3,25
Variance	0,866666667	0,6
Observations	16	16
df	15	
Stat t	2,070196678	
p-value	0,028054101	
t value	1,753050325	

Stat t > t value thus Reject H_0 Accept H_1

Question: Are oriented to business users.

	<i>Variable 1</i>	<i>Variable 2</i>
Mean	3,75	3,1875
Variance	0,866666667	1,095833333
Observations	16	16
df	15	
Stat t	1,781101839	
p-value	0,047575435	
t value	1,753050325	

Stat t > t value thus Reject H_0 Accept H_1

Question: Are oriented to Business Intelligence Activities.

	<i>Variable 1</i>	<i>Variable 2</i>
Mean	3,9375	3,375
Variance	0,729166667	0,516666667
Observations	16	16
df	15	
Stat t	2,33418733	
p-value	0,016951796	
t value	1,753050325	

Stat t > t value thus Reject H_0 Accept H_1

Question: Its full potential could be completely explored.

	<i>Variable 1</i>	<i>Variable 2</i>
Mean	3,6875	3,125
Variance	0,629166667	0,783333333
Observations	16	16
df	15	
Stat t	2,057534806	
p-value	0,028726682	
t value	1,753050325	

Stat t > t value thus Reject H_0 Accept H_1

Question: Help decision making.

	<i>Variable 1</i>	<i>Variable 2</i>
Mean	4,3125	3,75
Variance	0,4958333333	0,866666667
Observations	16	16
gl	15	
Stat t	2,182820625	
P(T<=t) uni-caudal	0,022678434	
t crítico uni-caudal	1,753050325	

Stat t > t value thus Reject H₀ Accept H₁

Question: Bring benefits to Higher Education Institutions.

	<i>Variable 1</i>	<i>Variable 2</i>
Mean	4,0625	3,5625
Variance	0,5958333333	0,529166667
Observations	16	16
gl	15	
Stat t	2,236067977	
P(T<=t) uni-caudal	0,020484478	
t crítico uni-caudal	1,753050325	

Stat t > t value thus Reject H₀ Accept H₁

Question: Bring benefits to organizations, in general.

	<i>Variable 1</i>	<i>Variable 2</i>
Mean	4,1875	3,75
Variance	0,5625	0,6
Observations	16	16
gl	15	
Stat t	2,405701888	
P(T<=t) uni-caudal	0,014747147	
t crítico uni-caudal	1,753050325	

Stat t > t value thus Reject H₀ Accept H₁

APPENDIX C – LIST OF PUBLISHED PAPERS

- Azevedo, A. & Santos, M.F. (2008). KDD, SEMMA and CRISP-DM: a Parallel Overview. In Weghorn, H. & Abraham, A. P. (Eds.), *Proceedings of the IADIS European Conference on Data Mining 2008*, IADIS Multi Conference on Computer Science and Information Systems, 182-185. Amsterdam, Holland: IADIS Press.
- Azevedo, A. & Santos, M.F. (2009). Business Intelligence: State of the Art, Trends, and Open Issues. In Liu, K. (Ed.), *Proceedings of the International Conference on Knowledge Management and Information Sharing (KMIS 2009)*, International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management (IC3K), 296-300. Funchal, Portugal: INSTICC.
- Azevedo, A. & Santos, M.F. (2009). An Architecture for an Effective Usage of Data Mining in Business Intelligence Systems. In Soliman, K. S. (Ed.), *Knowledge Management and Innovation in Advancing Economies: Analyses & Solutions*, Proceedings of 13th IBIMA Conference, 1319–1325. Marrakech, Morocco: IBIMA.
- Azevedo, A. & Santos, M.F. (2011). A Perspective on Data Mining Integration with Business Intelligence. In Kumar, A. (Ed.), *Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains* (pp.109-129). Hershey, NY: IGI Publishing.
- Azevedo, A. & Santos, M.F. (2011). QMBE: A Data Mining Language for Business Intelligence. *International Journal of Data Mining and Warehousing*, 1(1), 22-29.
- Azevedo, A & Santos, M. F. (2012). Binding Data Mining to Final Business Users of Business Intelligence Systems. Submitted to Conference.
- Azevedo, A & Santos, M. F. (2012). Closing the Gap Between Data Mining and Business Users of Business Intelligence Systems: a Design Science Approach. Submitted to Journal.

APPENDIX D –RIPPER ALGORITHM

The RIPPER algorithm description, as presented by the authors of the WEKA implementation is hereby included.

Used variables:

p = number of positive examples covered by the rule

n = number of negative examples covered by the rule

$t = p + n$

P = number of positive examples of the class

N = number of negative examples of the class

$T = P + N$

Algorithm:

Initialize the Rule Set: $RS = \{\}$, and for each class from the less prevalent one to the more frequent one, DO:

1. Building stage:

Repeat 1.1 and 1.2 until the Description Length (DL) of the rule set and examples is 64 bits greater than the smallest DL met so far, or there are no positive examples, or the error rate $\geq 50\%$.

1.1. Grow phase:

Grow one rule by greedily adding antecedents (or conditions) to the rule until the rule is perfect (i.e. 100% accurate). The procedure tries every possible value of each attribute and selects the condition with highest information gain: $p(\log(p/t) - \log(P/T))$.

1.2. Prune phase:

Incrementally prune each rule and allow the pruning of any final sequences of the antecedents;

The pruning metric is $(p-n)/(p+n)$ – but it's actually $2p/(p+n) - 1$, so in this implementation we simply use $p/(p+n)$ (actually $(p+1)/(p+n+2)$, thus if $p+n$ is 0, it's 0.5).

2. Optimization stage:

After generating the initial rule set $\{R_i\}$, generate and prune two variants of each rule R_i from randomized data using procedure 1.1 and 1.2. But one variant is generated from an empty rule while the other is generated by greedily adding antecedents to the original rule. Moreover, the pruning metric used here is $(TP+TN)/(P+N)$. Then the smallest possible DL for each variant and the original rule is computed. The

variant with the minimal DL is selected as the final representative of R_i in the ruleset. After all the rules in $\{R_i\}$ have been examined and if there are still residual positives, more rules are generated based on the residual positives using Building Stage again.

3. Delete the rules from the rule set that would increase the DL of the whole rule set if it were in it and add the resultant rule set to RS.

ENDDO

APPENDIX E – THE DATA MINING MODEL OBTAINED USING RIPPER

=== Model information ===

Filename: regras_JRip.model

Scheme: weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1

Relation: DadosParaDM-CSV-weka.filters.unsupervised.attribute.Remove-R2-3,7

Attributes: 20

Season ID

Course ID

Department

#of theoretic hours

#of theoretical-practical hours

Optional?

ECTS credits

Student Area

Student Nationality

Student Gender

Student Age

Student qualification

Student admission type

Attendance type

Program ID

Level description

Teacher rank

Teacher qualification

Teacher years on duty

Teacher age

=== Classifier model ===

JRIP rules:

=====

(Department = Informática) and (Program ID <= 3200) and (Season ID = N) and (Student Age >= 24) and (Student Age <= 26) => Level description=ND (5.0/1.0)

(Teacher years on duty <= 0) and (Optional? = S) and (Season ID = N) and (Student qualification = SEC) => Level description=ND (12.0/5.0)

(Department = Línguas e cultura) and (Student Age >= 29) and (Optional? = S) and (Teacher rank = Professor Adjunto) and (Student Age >= 34) => Level description=Muito Bom (30.0/13.0)

(Department = Línguas e cultura) and (Optional? = S) and (Teacher rank = Professor Adjunto) and (Student Area = AVEIRO) and (Course ID >= 2232) => Level description=Muito Bom (6.0/1.0)

(Season ID = N) and (Student Age >= 24) and (Attendance type = O) and (Student Age >= 27) and (Course ID >= 2370) and (Student Age >= 35) and (Student Gender = F) and (Course ID >= 2434) => Level description=Não Inscrito (24.0/5.0)

(Season ID = N) and (Student Age >= 24) and (Attendance type = O) and (Student Age >= 27) and (Student admission type = RI) and (Student Age <= 31) and (Student Area = GUARDA) => Level description=Não Inscrito (8.0/0.0)

(Season ID = N) and (Student Age >= 24) and (Attendance type = O) and (Student Age >= 27) and (Student admission type = RI) and (Student Gender = F) and (Program ID >= 3600) => Level description=Não Inscrito (8.0/0.0)

(Program ID >= 3200) and (Course ID >= 2218) and (Student Gender = F) and (Teacher rank = Professor Adjunto) and (Course ID <= 2439) and (ECTS credits >= 6) and (Student Age >= 28) => Level description=Bom (66.0/27.0)

(Program ID >= 3200) and (Course ID >= 2222) and (Student Gender = F) and (Teacher rank = Professor Adjunto) and (Program ID <= 4100) and (Teacher age >= 50) and (Student Age >= 33) => Level description=Bom (27.0/11.0)

(Program ID >= 3200) and (Course ID >= 2218) and (Season ID = N) and (Teacher age <= 42) and (Student Gender = F) and (ECTS credits <= 3) and (Teacher rank = Equiparado Assistente do 2º Triénio) and (Program ID >= 3300) => Level description=Bom (35.0/12.0)

(Program ID >= 3200) and (Course ID >= 2222) and (Student Gender = F) and (Season ID = N) and (Department = Gestão) and (Program ID >= 3800) and (Student Age >= 26) => Level description=Bom (27.0/11.0)

(Season ID = N) and (Department = Economia) and (Course ID <= 2352) and (ECTS credits <= 4) and (Student Age <= 24) => Level description=Sem nota mínima (740.0/287.0)

(Season ID = N) and (Course ID <= 2077) and (Teacher age <= 53) and (Course ID >= 1045) and (Course ID <= 1085) and (Student Age <= 22) and (Program ID >= 3600) => Level description=Sem nota mínima (471.0/159.0)

(Season ID = N) and (Department = Economia) and (Course ID <= 2352) and (ECTS credits <= 4) => Level description=Sem nota mínima (462.0/207.0)

(Season ID = N) and (Course ID <= 2077) and (Department = Economia) and (Course ID <= 1085) and (Attendance type = O) => Level description=Sem nota mínima (130.0/62.0)

(Season ID = N) and (Course ID <= 2077) and (Teacher age <= 53) and (Teacher years on duty >= 19) and (Teacher age <= 48) and (Teacher years on duty >= 22) => Level description=Sem nota mínima (229.0/113.0)

(Season ID = N) and (#of theoretical-practical hours >= 6) and (Course ID <= 2346) => Level description=Sem nota mínima (784.0/336.0)

(Season ID = N) and (Course ID <= 2105) and (Department = Direito) and (Course ID <= 1030) and (Student Gender = M) => Level description=Sem nota mínima (225.0/102.0)

(Season ID = N) and (Course ID <= 2105) and (Department = Economia) and (Attendance type = T) and (Teacher qualification = Mestre) => Level description=Sem nota mínima (182.0/73.0)

(Season ID = N) and (Department = Direito) and (Course ID <= 2313) and (Teacher years on duty >= 22) and (Attendance type = O) and (Course ID >= 2313) => Level description=Sem nota mínima (162.0/71.0)

(Season ID = N) and (Course ID <= 2105) and (Department = Direito) and (Course ID <= 1030) and (Attendance type = O) and (Student Age >= 20) and (Student Age <= 23) => Level description=Sem nota mínima (56.0/23.0)

(Season ID = N) and (Course ID <= 2105) and (Teacher age <= 53) and (ECTS credits >= 5) and (Course ID >= 2021) and (Student Age <= 20) and (Student Gender = M) => Level description=Sem nota mínima (63.0/24.0)

(Season ID = N) and (Course ID <= 2105) and (Course ID >= 2009) and (ECTS credits >= 5) and (Teacher qualification = Doutoramento) and (Course ID <= 2021) and (Student Age >= 20) and (Student Age <= 24) => Level description=Sem nota mínima (75.0/35.0)

(Department = Contabilidade) and (Season ID = N) and (Course ID <= 2350) and (Course ID >= 2348) => Level description=Faltou (1287.0/520.0)

(Student Age \geq 23) and (Department = Economia) and (ECTS credits \leq 4) and (Teacher rank = Equiparado Assistente do 2º Triénio) and (Attendance type = 0) and (Teacher age \geq 40) => Level description=Faltou (113.0/46.0)

(Student qualification = SEC) and (Student Age \geq 23) and (Course ID \leq 2369) and (Course ID \geq 2347) and (Season ID = N) and (Course ID \leq 2350) and (Student Age \geq 25) => Level description=Faltou (275.0/135.0)

(Student Age \geq 21) and (Department = Economia) and (ECTS credits \leq 4) and (Teacher years on duty \geq 22) and (Season ID = N) => Level description=Faltou (109.0/48.0)

(Student Age \geq 23) and (Student qualification = SEC) and (Course ID \leq 2369) and (Teacher age \leq 49) and (Program ID \leq 3200) and (Student Age \geq 26) and (Attendance type = 0) and (Student Age \geq 39) => Level description=Faltou (46.0/12.0)

(Student Gender = M) and (Student qualification = SEC) and (Course ID \leq 2369) and (Course ID \geq 2337) and (Teacher age \leq 49) and (Teacher age \geq 48) => Level description=Faltou (154.0/70.0)

(Season ID = R) and (Program ID \leq 3100) and (Course ID \leq 2347) and (Course ID \geq 2346) => Level description=Reprovado (581.0/171.0)

(Season ID = R) and (Program ID \leq 3100) and (Course ID \leq 2355) and (Department = Matemática) => Level description=Reprovado (292.0/80.0)

(Season ID = R) and (Course ID \leq 2355) and (Program ID \leq 3100) and (Student Age \leq 21) and (Student Area = BRAGA) => Level description=Reprovado (93.0/29.0)

(Season ID = R) and (Course ID \leq 2355) and (Program ID \leq 3100) and (Teacher age \geq 44) and (Teacher years on duty \leq 21) and (Course ID \leq 2348) and (Attendance type = T) and (Teacher age \leq 51) and (Student Age \leq 28) => Level description=Reprovado (64.0/21.0)

=> Level description=Suficiente (21322.0/14832.0)

Number of Rules : 34