Towards Real-World Data Mining

Sarabjot S. Anand[†], Alex G. Büchner[†], David A. Bell[‡], John G. Hughes[†]

[†] Northern Ireland Knowledge Engineering Laboratory

[‡] School of Information and Software Engineering
Faculty of Informatics, University of Ulster, Northern Ireland
E-mail: {ss.anand, ag.buchner, da.bell, jg.hughes}@ulst.ac.uk

Abstract

Data Mining or Knowledge Discovery in Databases has grown dramatically within research and industry during the last few years, and first projects have been finished successfully. But there are still plenty of open questions to be answered, and overall strategic concepts to be determined. To Propose a strategic business perspective on Data Mining is the major objective of this paper. It is also an attempt to lay down guidelines for the implementation of successful real-world Data Mining solutions and provide guidance to corporate business managers contemplating a Data Mining business solution.

We give a brief historical perspective of data management, which includes an account of the amount of data we have to date, its location, and a hypothetical but realistic look in the near future. We then describe the disciplines related to Data Mining such as database reporting and query tools, on-line analytical processing (OLAP), statistical data analysis tools and machine learning tools, and distinguish them from Data Mining. An overall Data Mining strategy is subsequently proposed, which shows the involved hidden costs and risks of undertaking such projects, using real-world

examples. The Data Mining process, is proposed next, and technologies which are required to fulfill tasks of the Data Mining process are described, before future directions are outlined.

1. Introduction

Over the past two decades there has been a huge increase in the amount of data being stored in databases as well as the number of database applications in business and the scientific domain. This explosion in the amount of electronically stored data was accelerated by the success of the relational model for storing data and the development and maturing of data retrieval and manipulation technologies. While technology for storing the data developed quickly to keep up with the demand, little stress was laid on developing software for analyzing the data, until recently when companies realised that hidden within these masses of data was a resource that was being ignored. The huge amount of stored data contains knowledge about a number of aspects of their business waiting to be harnessed and used for more effective business decision support. The Database Management Systems used to manage these data sets at present only allow the user to access information explicitly present in the

databases i.e. the data. The data stored in the database is only the tip of the 'iceberg of information' available from it. Contained implicitly within this data is knowledge about a number of aspects of business operations waiting to be harnessed and used for more effective business decision support. This extraction of knowledge from large data sets is called Data Mining[1] and is defined as the non-trivial extraction of implicit, previously unknown and potentially useful information from data [FRAW91]. The obvious benefits of Data Mining have resulted in a lot of resources being directed towards its development.

Almost in parallel with the developments in the database field, machine learning research was maturing with the development of a number of sophisticated techniques based on different models of human learning. Learning by example, cased-based reasoning, learning by observation and neural networks are some of the most popular learning techniques that were being used to move forward towards the goal of a "thinking machine".

While the main concern of database technologists was to find efficient ways of storing, retrieving and manipulating data, the main concern of the machine learning community was to develop techniques for learning knowledge from data. It soon became clear that what was required for Data Mining was a marriage between technologies developed in the database and machine learning communities. In fact, Data Mining can be considered to be an inter-disciplinary field involving concepts from Machine Learning, Database Technology, Statistics, Mathematics,

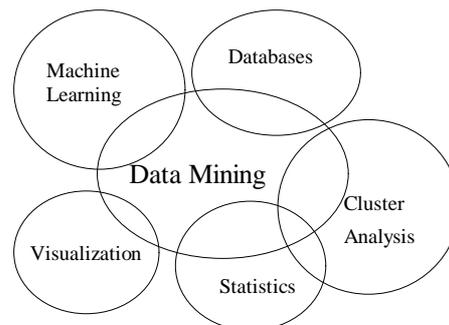Clustering and Visualisation among others (see Figure 1).

Figure 1: Data Mining

The need for Data Mining was felt due to the awareness that relational database management system (RDBMS) products were inappropriate for transforming data stored in on-line transaction processing (OLTP) applications into useful information and also to get "richer and deeper information from the information retrieval-oriented (IR) applications.. There are a number of specific reasons why RDBMSs are not suitable for business analysis [CODD93] including:

- To make OLTP queries fast, RDBMS applications generally normalise the data into 50 - 200 tables. Though great for OLTP operations, this is a nightmare for business analysis applications as it means a large number of joins are required to access the data necessary for such applications.
- While parallel processing can be useful in table scans it offers very little performance enhancement for complex joins.
- SQL is not designed with common business needs in mind. There is no way of using standard SQL for retrieving information like "the top 10 salespersons", "bottom 20% of customers", "products with a market share of

---

[1] This terminology does not conform with the terminology proposed by [KLOE96], but mirrors the common usage in industrial environments.

greater than 25%" or "the sales ratio of cola to root beer".

- RDBMSs do not provide common data analysis tools like data rotation, drill downs, dicing and slicing
- To allow truly ad-hoc end-user analysis, the database administrator should, ideally, index the database on every possible combination of columns and tables that the end user may ask for. This would create an unnecessary overhead for OLTP and query response times.
- Locking models, data consistency schemes, and caching algorithms are based on the RDBMS being used for OLTP applications where the transactions are small and discrete. Long running, complex queries cause problems in each of these areas.

As with any new technology transferred to the industry from the research laboratories, Data Mining does have a certain amount of hype associated with it and a number of vendors present it as the cure for all ills. Although such hype does produce short term gains, it can only damage the long term applicability of Data Mining. A prime example of a technology that suffered due to similar hype levels is expert systems. A rebirth of expert systems as knowledge-based systems has finally resulted in a less hyped but more successful technology. However, it delayed its application within the industry by almost a decade.

The objective of this paper is to present a business perspective on Data Mining. It is also an attempt to lay down guidelines for the implementation of successful real-world Data Mining solutions. It additionally provides guidance to corporate business managers contemplating a Data Mining business solution.

The outline of this paper is as follows: Section 2 gives a brief historical perspective of data management, which includes a justification of the amount of data we have to date, its location, and a hypothetical but realistic look in the near future. Section 3 describes the various disciplines related to Data Mining such as database reporting and query tools, on-line analytical processing, machine learning tools, statistical data analysis, and distinguish them from Data Mining. Next we propose an overall Data Mining strategy. Section 4 outlines the Data Mining process, describing the various stages of a Data Mining project within the real-world. In section 5 technologies which are required to fulfil tasks of the Data Mining process are described. Then in Section 6 we discuss some of the hidden costs and risks of involved in Data Mining projects, using real-world examples, before concluding the paper in Section 7.

2. A Historical Perspective of Data Management

Data management started about three decades ago when data was stored in flat ASCII or EBCDIC operating system files without any information about the data. Often data had to be stored more than once across the organisation leading to inconsistencies and inefficiencies. There were no query languages, and any constraints or interrelationships among the entities was left to applications accessing the files. Database Management Systems were introduced in the late 1960's largely triggered by the Space Race. They provided these missing facilities and were based on either networks, hierarchies or (later) relations. Constraints, such as data types, value ranges, dependencies, or relationships among entities were provided and meta data stored in data

dictionaries. Later still, fourth generation languages were provided to ease application development. Within the last decade these systems have been extended to handle distributed and heterogeneous data, and thus more semantics about the inter-relationships of different sites had to be embedded. Richer modelling methods such as the Semantic data model and Object-oriented data models have now been developed and are making their mark in the industry.

State-of-the-art data management systems are platform-independent client-server solutions with active back-ends and visualised front-ends. In addition to user friendly manual data entering, these components allow automatic data generation, which leads to an enormous data volume increase (as [FRAW91] stated "the amount of data doubles every 20 months"). The information hidden in both, the manually entered and the automatically generated data, is tremendous.

So, what have we been doing with all that data to date, and what will we do in the near future? Surely, the data flood won't stop, but we can slow down the increase in storage space needed, using knowledge based pre-processing techniques. That is, we will delete data, but not the information stored in the data. Knowledge discovering techniques will make — have already made — a major impact on how we view our data, and the decisions we draw from that knowledge. That is, in addition to transaction and analysis orientated data processing, there will be a third component, which is to do with processing knowledge about the data. This solves quite a few problems outlined above, but also introduces new obstacles:

- How do we know that the knowledge is correct?
- How can we manage knowledge in a similar way we handle data?
- What about data privacy aspects?

These questions, among others, have to be answered in the near future to give Data Mining the respect it deserves in industry and research.

Data Warehousing is currently a "hot topic" in data management. It is a technique for integrating "legacy" systems within a corporation to provide an enterprise wide view to decision making. This technology has become necessary due to the realisation on the part of large organisations that decisions about one business process cannot be made in complete isolation of other business processes within the enterprise. For example, large financial organisations may have different sections of their marketing departments maintaining their own customer data based around different products. The individual product-centred customer databases often do not link in together. While such a situation may be fine from a day-to-day operational perspective clearly, from a decision support perspective a much more beneficial situation would be a customer-centred customer database where the same customer identification number is used to identify all the different products bought by each customer.

Also, most large corporations have operational data in production systems that is unreliable and disparate, making it difficult to integrate or extract for analysis purposes. Thus the implementation of a Data Warehouse consists of the acquisition of data from multiple internal and external sources (of the corporation), the management and integration into a central, integrated repository, the

provision of access, reporting and analysis tools to interpret selected data converting it into information to support managerial decision making processes.

However, in a recent survey of Data Warehousing projects ([CONS96]) it was reported that while most warehousing projects meet the users requirements of data quality and integration the ability to analyse and convert the data into information has been a major disappointment. Data Mining and other data analysis tools attempt to provide this ability of converting data into information to the Data Warehouse user. According to Charles Bonomo, vice-president of Advanced Technologies at J. P. Morgan, a large financial organisation in the US, "One of the primary justifications for implementing a data warehousing solution is having a Data Mining tool in place that can access the data within it". The most logical consequence is to integrate Data Mining technology as central part of the data Warehousing philosophy, without limiting any of the existing functionality.

3. Related Disciplines

There are a number of technologies that are related to Data Mining. These include database query and reporting tools, on-line analytical processing (OLAP), statistical data analysis and machine learning tools. In this section we describe each of these disciplines and compare and contrast their objectives with those of Data Mining.

Data Mining has been defined as the "tireless and relentless searching of data for useful patterns" [BYTE]. This tireless and relentless searching can either be automatic or manually guided. Manually guided systems have the disadvantage that the type of patterns searched for, needs to be conceived by the user and less obvious patterns remain undiscovered. Query and reporting tools, as well as OLAP, fall into the category of manually guided systems, whereas machine learning and Data Mining tools are automated discovery tools although allowing for user biases.

We consider Data Mining to be automated searching systems that still require a certain amount of input from the user but do not need to be directed by the user as to what patterns have to be discovered, sometimes referred to as "semi-automisation". We now distinguish the above mentioned related technologies from Data Mining and describe the inter-relationships amongst them.

3.1. Database Reporting Tools

Cynics have often accused Data Mining of being a new buzz world for Database Management System (DBMS) Reports. However, this is clearly not the case. Using a DBMS Reporting tool a company could generate reports such a "Last months sales for each service type", "Sales per service grouped by customer sex or age bracket" and "List of customers who lapsed their insurance policy". Data Mining allows the user to ask less specific questions that cannot be asked using a DBMS Reporting tool. Data Mining supports questions such as - "What characteristics do my customers that lapse their policy have in common and how do they differ from my customers who renew their policy?" and "Which of my motor insurance policy holders would be potential customers for my House Content Insurance policy?".

It also can be observed that modern Reporting tools embody more and more data analysis related facilities, such as drill-down

reports, dicing, slicing or rotating. Additionally, statistical functionality including graphical interpretation of results has become a standard feature of each report generator. So perhaps the gap between these types of tools will reduce over time.

3.2. Statistical Data Analysis

Statistics falls short of the goals of Data Mining. Firstly, Statistics is ill-suited for nomial and structured data types that are common in real-world databases. Secondly, Statistics is totally data driven and does not provide techniques for incorporating domain or prior knowledge. Thirdly, the process of statistical data analysis requires expert user guidance. Lastly, the results from a statistical analysis are difficult to interpret and are overwhelming to non-statisticians.

As with Query Reporting tools, modern statistical packages extend the providence of tools and techniques towards Data Mining. Examples are open database interfaces, parallelism support, or embodying of basic machine learning techniques e.g. SAS includes Neural Network software as one of its modules.

3.3. On-Line Analytical Processing

Complex statistical functionality was never intended to be accommodated within RDBMSs. Providing such functionality was left to user-friendly end-user products such as spreadsheets or statistical packages which are supposed to act as front ends to the RDBMS. Though statistics packages and related tools have provided a certain amount of functionality required by business analysts, none address, to any great extent, the need for analysing the data according to its multiple dimensions. Any product that intends to provide such functionality to business analysts must provide

the following features to allow adequate statistical data analysis:

- Access to many different types of files
- Creation of multi-dimensional views of the data
- Experimentation with various data formats and aggregations
- Definition and visual animation of new information models
- Application of summations and other formulae to these models
- Drilling down, rolling up, slicing and dicing, rotation of consolidation paths
- Generation of a wide variety of reports, charts and diagrams

On-line Analytical Processing (OLAP) is the name given by E. F. Codd ([CODD93]) to the technologies that attempt to address these user requirements. Codd defined OLAP as "the dynamic synthesis, analysis and consolidation of large volumes of multi-dimensional data". Codd provided 12 rules/ requirements of any OLAP system. These are:

- Multi-dimensional conceptual view
- Transparency
- Accessibility
- Consistent reporting performance
- Client-server architecture
- Generic dimensionality
- Dynamic sparse matrix handling
- Multi user support
- Unrestricted cross-dimensional operations
- Intuitive data manipulation
- Flexible reporting
- Unlimited dimensions and aggregation levels

A number of extensions to these OLAP requirements have been suggested since then [DRES93, BUYT95, MENN94], including: Support for multiple arrays, time series analysis, OLAP joins, procedural language and

development tools, database management tools, object storage, integration of functionality, subset selection, detail drill down, local data support, incremental database refresh and an SQL interface.

Data is stored in an n-dimensional hypercube based on different pre-defined business metrics. This hypercube is then viewed from any chosen two dimensions. OLAP provides statistical information along the different dimensions and is very efficient for up to 15 dimensions.

3.4. Machine Learning Tools

The main factor that distinguishes Data Mining from machine learning is that it is about learning from existing real-world data rather than data generated particularly for the learning tasks. In Data Mining the data sets are large therefore efficiency and scalability of algorithms is important. As mentioned earlier the data from which Data Mining algorithms learn knowledge is already existing real-world data. Therefore, typically the data contains plenty of missing values as well as noise and it is not static i.e. it is prone to updates. However, as the data is stored in databases, efficient methods for data retrieval are available that can be used to make the algorithms more efficient. Also, domain knowledge in the form of integrity constraints is available that can be used to constrain the learning algorithms search space.

In summary, machine learning algorithms form the basis for most Data Mining tools. However, to make them suitable to handle real-world Data Mining problems appropriate extensions have to added to these techniques.

4. The Data Mining Process

Data Mining should be viewed as a *process* involving several automated and non-automated steps rather than a single step. In this section we describe each of the steps in the Data Mining process and discuss the different aspects of each of these steps.

4.1. Human Resource Identification

After a problem has been identified at the management level of an organisation *Human Resource Identification* is the first phase of the Data Mining process. In most real-world Data Mining problems the Human Resources required are: the Domain Expert, the Data Expert and the Data Mining Expert. Normally, Data Mining is carried out in large organisations where the prospect of finding a Domain Expert who is also an expert in the data stored by the organisation is rare. For example, in a large bank, the Domain Expert would belong to the sales department while the Data Expert will probably belong to the IT department. The Data Mining expert would normally belong to an organisation outside the bank deployed by the bank for the purpose of solving the Data Mining problem. We want to stress the need for bringing together these human resources early into the process as any project that does not bring together these expertise right at the beginning of the process will very likely encounter problems later on.

4.2. Problem Specification

Problem Specification is the second phase of the Data Mining Process. Here a better understanding of the problem is developed by the human resources identified in the Human Resource Identification component of the project. The problem is identified as a particular type of Data Mining task. A number of Mining tasks have been addressed in the literature. For example, discovery of association rules [AGRA94], classification rules [AGRA92], sequence rules [AGRA95] and characteristic

and discriminant rules [CAI91]. Different techniques are used to tackle each of these tasks. Therefore, identifying the task type that the problem falls into is important. Note that we are taking for granted here that the user has some idea of the problem he or she is trying to solve. In most cases this is a fair assumption as 'pure discovery', completely data driven, is rare.

The second part of the Problem Specification phase is to identify the ultimate user of the knowledge. Clearly if the knowledge discovered is to be used by a human, it must be in a format that the user can understand and is used to. However, if Data Mining is only a small part of a larger project and the output from Data Mining is to be interpreted by a computerised system, the format of the discovered knowledge will have to strictly adhere to the expected format.

4.3. Data Prospecting

Data held by corporations wanting to undertake a Data Mining exercise is the most important factor that needs to be taken into account when considering a Data Mining solution. If the data is not of a high quality knowledge extracted from the data cannot be expected to be of high quality. The old adage of "Garbage In, Garbage Out" still holds. The main problem with data in the real-world is that though databases are usually well designed (i.e. are a rich model of the real-world), the *population* of the individual fields within the database is low. The reason for this is simple. Since the early 1980s a number of companies have started collecting data electronically. This move towards data collection has caused a number of changes within the workplace which have been met with varied acceptance. For example, in an insurance company whenever a new policy is issued, the insurance agent normally fills in the minimal amount of data into the database as he/ she sees this as an unnecessary waste of time. The reason for this is that to date there has been no obvious benefit of data collection. It is important to educate the work force within the organisation to the benefits of data collection. This is an additional, hidden cost often overlooked by the corporate executives. In a recent pilot project undertaken with a large financial organisation in Ireland we found that some of the fields within their customer database had just 5% of their records filled in. Clearly, using such a field within the discovery could very easily skew the whole learning process given the unsatisfactory techniques available for dealing with missing values.

Another major problem encountered in the real world related to the available data in an organisation is the accessibility of the data. Most data within an organisation is stored in distributed, legacy systems making integration of the databases difficult. A number of Data Mining vendors dismiss this problem as being one that needs to be dealt with by data warehousing vendors. However, by doing so we are limiting the applicability of Data Mining technology. A number of traditionally low investors in technologies, for example, charities, have a need for Data Mining but cannot justify undertaking the expense of the implementation of a Data Warehouse. A good real-world example of such an organisation is the Craigavon Urological Research and Education (C.U.R.E) Charity based at Craigavon Area Hospital (CAH) in Northern Ireland. The present computerised systems, the Patient Audit System (PAS), Laboratory Regional Standard System (RSS), and the Radiology System at the CAH store and provide access to information required for day to day

administration of the hospital but do not contain clinically relevant information. The Urology consultant wanted to store clinically relevant information to perform intelligent audit of his patients leading to Urological research and better patient care. Some of the data required for such an Intelligent Clinical Audit System (InCAS) is already stored electronically within the PAS and RSS systems. However, these systems were implemented in isolation and do not contain a common key that can be used to integrate the data from the two systems. Clearly, a Data Warehouse is not a viable option for C.U.R.E. The integration of the PAS and RSS data is being undertaken manually.

Data Prospecting is therefore an important next phase in the process. It consists of analysing what is the state of the data required for solving the problem at hand. There are three main considerations within this phase: What are the relevant attributes? Is the data required electronically stored and accessible? Are the data attributes required populated?

4.4. Choosing the right Data Mining Methodology

The main task of the *Methodology Identification* phase is to find the best Knowledge Discovery methodology to solve the specified Mining problem. Often a combination of methodologies is required to solve the problem at hand. For example, clustering or data partitioning may be required before the application of a classification algorithm. The most commonly used technologies are rule induction, derivatives of traditional statistics, genetic algorithms, evidence theory, case-base reasoning, Bayesian belief networks, fuzzy logic, rough sets and neural networks. The chosen paradigm depends on the type of information that is required and the domain of knowledge being discovered. For example, if an explanation of the discovered knowledge is required neural networks would clearly not be an appropriate methodology. The selected technique also influences the format of the input data, whose preparation is part of the Data Pre-processing phase of the Data Mining process.

In certain cases it is difficult to find out which tools are most appropriate for a particular task. For example, most financial organisations want to employ Data Mining for cross-sales purposes. At first glance cross-sales seems like a simple classification problem. However, on closer examination it is clear that what we are dealing with is subtly different from a classification problem. We discuss the methodology requirements for the cross-sales problem below.

Consider the following scenario. A Bank wants to promote a product A by targeting existing customers who are not already customers of product A. Within their customer database, the bank has two types of customers:

Type 1: Those that have product A
Type 2: Those that do not have product A

For a classification problem what is required is three types of customers:

Type 1: Those that have product A
Type 2: Those that have refused to purchase product A
Type 3: Those that do not have product A but have not refused to purchase product A

The first type of customers form the positive examples, the second type of customers form the negative examples and the third type form the target data set.

Therefore, for cross-sales we only have positive examples and a target data set. Thus, rather than using a classification algorithm we

need to use an association algorithm to discover characteristic rules. These rules define those characteristics that are prevalent in a particular group of records which in our case is the group of records that pertain to customers who have product A. Given these rules customers in the target data set with similar characteristics can be targeted with sales campaigns. Characteristic rules are not as accurate as classification rules would be as classification rules take both the negative and positive examples into account when discovering the rules but, in general, they are more accurate than totally random selections.

An important aspect of using association algorithms for discovery purposes is the aspect of discovering only the interesting rules i.e developing a knowledge filter that would help reduce the volume of knowledge being discovered [ANAN96].

The next stage in the process of cross-sales would be the exploitation of the discovered rules by the targeting of the customers, from the target data set, picked using the characteristic rules. At this stage the bank can keep a record of those customers that were targeted but did not buy the product and use these records to refine the characteristic rules making them more accurate or they can use these records as negative examples and then use a classification algorithm to discover classification rules.
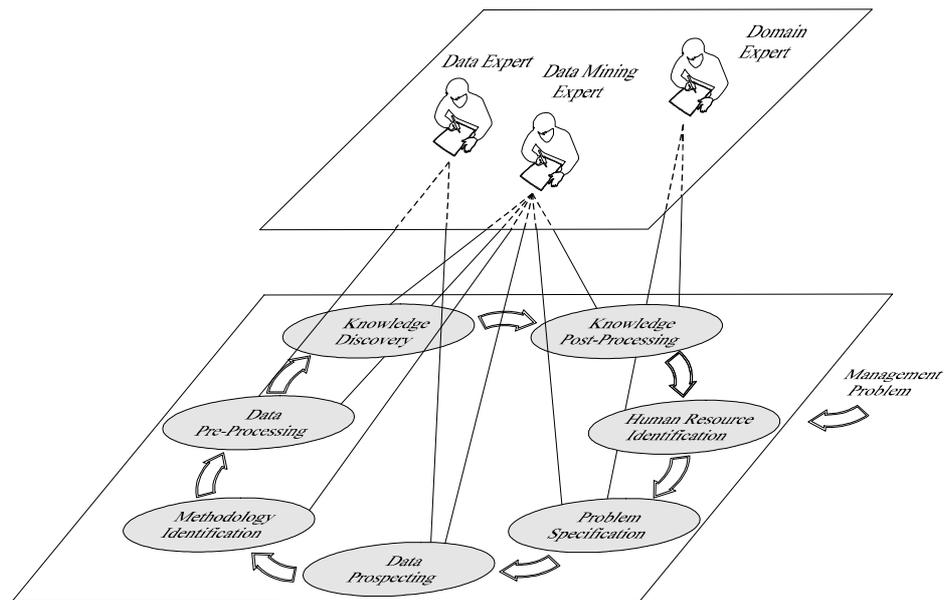
Clearly, if a detailed analysis of the problem

Figure 2: The Data Mining Process

at hand had not been carried out at the Problem Specification stage of the process it could lead to the incorrect methodology being chosen, resulting in the failure of the project.

4.5. Data Pre-Processing

The next phase is that of *Data Pre-processing*. Depending on the state of the data this process may or may not constitute the phase where most of the effort of the Data Mining Process is concentrated. Data Pre-processing involves removing outliers in the data, predicting and filling-in missing values [LITT87, QUIN86], noise modelling [HICK96], data dimensionality reduction using feature subset selection [KOHA95, CARU94], data quantisation, transformation and coding and heterogeneity resolution. Data Pre-processing in Data Mining has often been considered the equivalent of the 'stick of dynamite' within the mining of ores [BYTE]. No pre-processing will result in the

really useful knowledge remaining undiscovered. However, too much pre-processing may result in the discarding of the data in which the interesting knowledge is implicit. Thus, there is a balance required during pre-processing.

An important aspect of Data Pre-processing is the preparation of the data into a form that is acceptable to the chosen Data Mining methodology. For example, most association rules can discover knowledge only from single relations, thus if the data identified in the data prospecting stage exists in multiple relations within the database, the data needs to be pooled together before discovery can take place. Such a "data pooling" could require anything from a simple join of the relations to the creation of summary attributes that reflect 1-to-n relationships between the database relations. Clearly, the role of the data and domain experts in this data preparation would vary based on the approach taken.

4.6. Knowledge Discovery

The *Knowledge Discovery*[2] or *Pattern Discovery* phase follows the Data Pre-processing phase. It consists of using algorithms that automatically discover patterns from the pre-processed data. The choice of algorithm depends on the mining task at hand. Due to the large amounts of data from which knowledge is to be discovered, the algorithms used in this phase need to be efficient. Techniques for learning with sampling and high performance computing are important considerations within this phase.

From our experience the Knowledge Discovery task is ideally not totally automated and independent of user intervention. The Domain Expert can often provide domain knowledge that can be used by the discovery algorithm for making patterns in the data more visible, pruning of the search space or for filtering the discovered knowledge based on a user driven interest measure. This phase is often iterative with knowledge discovered after each iteration leading to the refinement of the discovery goal as well as the domain knowledge.

We would like to stress, here, upon the importance of domain knowledge refinement as the use of inaccurate domain knowledge may result in the loss of information and useful knowledge remaining undiscovered. [ANAN96] discusses the effect of domain knowledge on the quality of discovered knowledge in greater detail.

4.7. Knowledge Post-Processing

The last step of the Data Mining process is *Knowledge Post-processing*. Trivial and obsolete information has to be filtered out and discovered knowledge has to be presented in a user-readable way, using either visualisation techniques or natural language like constructs. Often the knowledge filtering process is domain as well as user dependent [ANAN96], and thus requires domain specific semi-automated techniques.

In the case of Association algorithm based Data Mining, the requirements of knowledge filtering and its effectiveness play an important role in the success of the project. This is mainly due to the enormous quantity of rules that can potentially be discovered using this discovery methodology. It has been our experience that

---

[2] This phase is referred to as Data Mining in [KLOE96], however, we use Data Mining and Knowledge Discovery in Databases interchangeably as it reflects the industrial use of the term Data Mining as the overall process.

the domain experts loose interest if they have to manually sift through a large number of rules.

4.8. Knowledge Maintenance

Due to the fact that the data used as input to the Data Mining process is often dynamic and prone to updates, the discovered knowledge has to be maintained. *Knowledge Maintenance* may consist of re-applying the already set up Data Mining process for the particular problem or using an incremental methodology that would update the knowledge as the data changes keeping them consistent.

The phases of the Data Mining Process outlined above (Figure 1) require technological support for the following tasks: Interfaces for the Domain Expert and Data Mining Expert, Access to Data, Data Pre-processing, Knowledge Discovery, and Knowledge Post-processing facilities. These are described briefly in the following section.

5. Technological Support

The phases of the Data Mining process outlined above (Figure 2) require technological support for a number of tasks. We now identify these tasks and outline the technical support required for these tasks to be successfully executed.

5.1 Interfaces for the Human Expert

Technological supports to allow the Human experts to interact with the system is required. These include interfaces for describing the data view for the discovery, describing syntactic constraints, support constraints, domain knowledge, interestingness measures, semantic equivalence relationships for heterogeneous data and user biases.

5.2 Access to Data

Flexible access to data is imperative. The discovery algorithms must be able to access distributed and heterogeneous data sources. Facilities for handling semantic heterogeneity are required.

5.3 Data Pre-processing Technologies

Facilities like interactive graphics for data selection, principal component analysis or factor analysis or feature subset selection [KOHA95] for data dimensionality reduction, thresholding for removal of outliers, statistical models for handling noise in the data [HICK96], Bayesian techniques for filling in missing values [LITT87, QUIN86]], information theoretic measures for data discretisation and semantic equivalence relationship handling are some of the technological support that may be provided for Data Pre-processing. Facilities are also required for exploratory data analysis that allow the experts to explore and formulate the problem data space.

5.4 Algorithmic Support

Discovery of different types of knowledge e.g. classification, association, characteristic, temporal, cluster analysis and discriminant, should be supported with different underlying uncertainty models so that one that is most appropriate to the problem at hand can be chosen. Facilities for using high performance computing and learning by sampling may be required for efficiency of discovery. The hardware support for the algorithm should be transparent to the user i.e. the algorithms should automatically be able to take advantage of any additional computing power made available to them.

5.5 Knowledge Post-processing Technologies

Flexible techniques for filtering out knowledge discovered so that obvious and uninteresting rules are not presented to the user are required. Graphical and other user oriented techniques for

displaying the knowledge are also required. An inference engine that allows the user to make use of the discovered knowledge effectively as well as techniques for verifying the discovered knowledge are also required.

6. Hidden Costs of Data Mining

Data Mining must be user-led, keeping the goals and objectives of the corporation in mind. Most IT-led projects do not succeed. Also, a corporation has a number of departments within it with varying business requirements and so no one tool can suffice. It is very important to gauge what kind of information is required by the corporate decision makers and tools appropriate for the particular requirement needs to be chosen. Most successful Data Mining projects have been where the tool chosen is specific to the market sector e.g. AT&T's Sales and Marketing Packs and GTEs Health-KEFIR in Health.

6.1 Effect on production systems

Production systems have been built with the collection and manipulation of data in mind. Their design is based around making specific On-line transaction processes (OLTP) efficient. When using data for decision support the requirements from the system are different. Data required needs to be collated from different database tables using table joins. Such joins create a large overhead on the production systems especially when they are distributed over more than 5-10 tables, which are commonplace in decision support queries. The OLTP type queries are therefore bound to suffer. Thus, in situations where the OLTP queries are crucial Data Mining exercises should be performed on secondary data stores e.g. Data Warehouses or "Backup" Databases, where possible.

6.2 Data Extraction Costs

In situations in which a Data Warehouse or secondary data store does not exist, data must be downloaded from production systems into flat ASCII files and Data Mining performed on a separate platform than those used by OLTP applications. This extraction of data is normally not straightforward either and depending on the Data Mining tool to be used, it may require some data transformation as well. Additional resources for this purpose must be taken into account when costing a Data Mining solution.

6.3 Resistance to Change - "The Law of Inertia"

The introduction of IT solutions within an organisation requires a whole new work ethic. IT affects employees at all levels of the organisation. Therefore, the introduction of IT is often met with resistance. We have already pointed out, in Section 4.3, how employees need to be made aware at the data acquisition stage of the process of benefits of the system. At the knowledge utilisation stage similar awareness exercises are necessary for the system to be successful as well as worthwhile. The Financial Times reported on the 28th of November 1995 in an article on Data Mining that "Marketing experts are often torn between admiration for analytical power that these technologies (Data Mining) provide, and regret as it is displacing creativity, intuition and judgement". However, this seems unjust as Data Mining only provides the decision maker with information that is needed to make the decision - what decision is reached at by using this information is still in the hands of decision maker. Thus, Data Mining empowers the marketing experts with information that they require to make better decisions - they do not hamper the decision process in any way.

7. Conclusions

In this paper we have discussed electronic data recording and the lack of analysis of the data that has been carried out to date in most commercial, data rich sectors. We have discussed techniques that are available at present for data analysis including query reporting tools, statistical data analysis tools, on-line analytical processing tools and machine learning tools, and compared and contrasted them with the aims of technologies that have come to be known as Data Mining. We have attempted to put together an overall Data Mining strategy that included a discussion on the various stages or phases of the Data Mining process, the human factor in the process, the technological requirements of the process and the hidden costs. The objective has been to provide guidelines to people in the industry who are contemplating an Data Mining solution to their decision support problems.

Though a lot of effort is being put into Data Mining in academia and industry and a number of successes in the real-world are being reported [FAYY96], a number of hard questions still remain unanswered. Most of the Data Mining solutions revolve around relational data stores. However, a large part of the electronically stored data is unstructured, multimedia data. How is knowledge to be discovered from such unstructured, distributed and heterogeneous data sources? Knowledge filtering techniques are still in their infancy and require considerable effort from the domain and Data Mining expert. Learning in noisy and incomplete data is still difficult. In a number of cases the data used for learning a concept is too sparse making learning difficult. Discovery in dynamic environments has not been addressed.

8. References

[AGRA92] R. Agrawal, S. Ghosh, T. Imielinski, B. Iyer and A. Swami, An interval classifier for database mining applications, Proc of 18th Int. Conf. on VLDB, Pg. 560-573, 1992.

[AGRA94] R. Agrawal, R. Srikant, Fast Algorithms for Mining Association Rules in Large Databases, Proc. of the 20th VLDB Conference, Pg. 487 - 499, Chile, 1994

[AGRA95] R. Agrawal, K. Lin, H. S. Sawhney, K. Shim, Fast Similarity Search in the Presence of Noise, Scaling and Translation in Time-Series Databases, Proc. of the 21st VLDB Conference, 1995.

[ANAN96] S.S. Anand, J.G. Hughes, D.A. Bell, A. Patrick, Tackling the Cross-Sales Problem using Data Mining, Submitted for publication, 1996.

[BYTE] DeJeasus E X, Hedberg S R, Watterson K, Krivda C D (1995). Data Mining, BYTE October Issue.

[BUYT95] F A. Buytendijk OLAP: playing for keeps (maintenance and control aspects of OLAP applications), July 1995

[CAI91] Y. Cai, N. Cerecone, J. Han, Attribute-Oriented Induction in Relational Databases, Knowledge Discovery in Databases. Pg. 213 - 228. AAAI/ MIT Press, 1991.

[CARU94] R. Caruana, D. Freitag. Greedy Attribute Selection, Proc. of the 11th Int. Conf. on Machine Learning, Morgan Kaufmann, Pg. 28 - 36, 1994.

[CODD93] Codd E F, Codd S B and Salley C T (1993). Providing OLAP to User-

Analysts: An IT Mandate, White Paper produced by Codd and Date Inc.

[DRES93] Dresner, OLAP: Heightened Industry Focus on Business Intelligence, Gartner Group, October 4, 1993

[FAYY96] Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy P (1996). Advances in Knowledge Discovery and Data Mining, AAAI/ MIT Press, 1996.

[FRAW91] W. J. Frawley, G. Piatetsky-Shapiro, C. J. Matheus, Knowledge Discovery in Databases : An Overview Knowledge Discovery in Databases, Pg. 1 - 27, AAAI/MIT Press 1991.

[HICK96] R. J. Hickey, Noise Modelling and Evaluating Learning from Examples, Artificial Intelligence, Vol. 80, 1996.

[KLOE96] W. Kloesgen, J.M. Zytkow, Knowledge Discovery in Databases Terminology, in Advances in Knowledge Discovery and Data Mining, U.M. Fayyad, et. Al.(eds.), Pg.. 573-592, 1996.

[KOHA95] R. Kohavi, D. Sommerfield. Feature Subset Selection Using the Wrapper Method: Overfitting and Dynamic Search Space Topology, Proc. of the 1st Int. Conf. on Knowledge Discovery and Data Mining (KDD-95), 1995.

[LITT87] R. J. A. Little, D.B. Ruthlin. Statistical Analysis with Missing Values, Wiley, 1987.

[MENN94] D. Menninger, OLAP Turning corporate data into Business Intelligence, IRI Software, 1994.

[QUIN86] J. R. Quinlan, Induction of Decision Tree , Machine Learning 1, Pg. 81-106, 1986.