DOWNLOAD PDF OVERVIEW OF THE DATA MINING PROCESS

Chapter 1 : Data Mining Overview

By: Siddharth Mehta Overview. Data Mining can be applied for a variety of purposes. Before one starts considering data mining as a probable solution, one should clearly understand the typical applications of data mining as well as the approach to develop data mining models in an enterprise.

Anand and John G. What is Data Mining? 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 fast to keep up with the demand, little stress was paid to developing software for analysing the data until recently when companies realised that hidden within these masses of data was a resource that was being ignored. The huge amounts 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. Database Management Systems used to manage these data sets at present only allow the user to access information explicitly present in the databases i. Contained implicitly within this data is knowledge about a number of aspects of their business waiting to be harnessed and used for more effective business decision support. This extraction of knowledge from large data sets is called Data Mining or Knowledge Discovery in Databases 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 has 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 create the ultimate thinking machine. Data Mining 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. 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. So how does Data Mining differ from Machine Learning? After all the goal of both technologies is learning from data. Data Mining 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 lots of missing values and noise and it is not static i. 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. This is not true. Last months sales for each service type Sales per service grouped by customer sex or age bracket List of customers who lapsed their insurance policy However, using Data Mining techniques the following questions may be answered What characteristics do my customers that lapse their policy have in common and how do they differ from my customers who renew their policy? Which of my motor insurance policy holders would be potential customers for my House Content Insurance policy? Characteristics of a Potential Customer for Data Mining Most of the challenges faced by data miners stem from the fact that data stored in real-world databases was not collected with discovery as the main objective. Storage, retrieval and manipulation of the data were the main objectives of the data being stored in databases. Thus most companies interested in data mining poses data with the following typical characteristics: The stored data is large and noisy Conventional methods of data analysis are not useful due to the complexity of the data structures and the size of the data The data is distributed and heterogeneous due to most of the data being collected over time in

legacy systems The sheer size of the databases in real-world applications causes efficiency problems. The noise in the data and heterogeneity cause problems in terms of accuracy of the discovered knowledge and complexity of the discovery algorithms required. Aspects of Data Mining In this section we discuss a number of issues that need to be addressed by any serious data mining package. Nothing is certain in this world and therefore any system that tries to model a real-world scenario must allow a representation for uncertainty. A number of uncertainty models have been proposed in the Artificial Intelligence community. Though no consensus has been arrived at as to which model is best it is recognised that attention must be paid on the selection of a model that is suitable for the problem at hand. Dealing with Missing Values: Missing Values can occur in databases due to two reasons: Firstly, a value may not be available at the present time incomplete information and secondly, no value may be appropriate due to some other attributes value in the tuple. Within the relational model missing values are represented as NULLs. Facilities must be provided to deal with NULL values within a Data Mining system either by filling in these values before the discovery process is undertaken or by taking NULLs into account within the discovery process, perhaps by using a model of uncertainty like Evidence Theory that allows an explicit expression for ignorance. A number of methodologies have been suggested in machine learning literature e. NULL as an attribute value, using the most common attribute value and decision tree techniques. Dealing with Noisy data: Noise in data in real-world databases are a fact of life. Discovery techniques used for Data Mining therefore need to be able to handle noisy data. As compared to symbolic learning techniques like decision tree induction, Neural Network techniques tend to generalise and learn classification knowledge better in the presence of knowledge. Though a number of techniques based on statistics have been used in machine learning techniques more robust techniques are required in Data Mining for dealing with noise if useful discovery from data is to be performed. Machine Learning algorithms though highly sophisticated and general get very inefficient when used for learning from large data sets. In Data Mining the data sets are very large and therefore the need to create new efficient, more specific algorithms is very important. From large amounts data an even larger amount of knowledge can be discovered. Therefore what is required is techniques that prioritise the knowledge in terms of its usefulness or interesting to the present needs of the user. At present the uncertainty and support of the knowledge, knowledge about the user domain and some measure of interestingness is used. The measure of interestingness is accepted as being a subjective measure as what is interesting to one user may be of no interest to another. However, some aspects of interestingness can be automated and a number of measures have been suggested e. Very often some reliable knowledge about he discovery domain may be available to the user. An important question is how to use this knowledge to discover better knowledge in a more efficient way. Size and Complexity of Data: As compared to machine learning problems the data sets in Data Mining are much larger, noisier and incomplete. Also the data used for discovering knowledge in Data Mining was not collected or stored for the purpose of discovery. Most data has been collected over a period of time and lies in different formats in legacy systems. Thus, heterogeneity and distribution of data is of particular interest to Data Mining. Techniques are required for integrating heterogeneous and distributed data. Due to the large amounts of data, efficiency of the Data Mining algorithms is important. One way of improving the efficiency of Data Mining techniques is by reducing the amount of data. A lot of work has been done in Machine Learning with respect to relevance. Similar techniques need to be employed in Data Mining. Understandability of Discovered Knowledge: Knowledge discovered using Data Mining techniques must be in a form that can be understood by the user as in the end of the day a user will only be able to use the knowledge for decision making if he or she is able to understand the knowledge. This is the main failing of Neural Networks as they are unintelligible black boxes. Decision Trees can get very large and opaque when using a large training data set. Consistency between Data and Discovered Knowledge: Data stored in databases may be updated from time to time. Techniques are required for updating the knowledge discovered from the data so that it is consistent with updates made to the data. Classification of Data Mining Problems Agrawal et. Consider a bank that gives loans to its customers. The bank would obviously find it useful to be able to predict which new customer would be a good investment and which one would not. Using data collected about the previous customers, the bank would like to know the attributes that make a customer a good investment or a bad investment. What is required is a set of rules that

partition the data into two exclusive groups - one of good investments and the other of bad investments. Such rules are called classification rules as they classify the given data into a fixed number of groups. The data on old customers for whom the group that they belong to is known is called the training set from which the classification rules are discovered. The classification rules can then be used to discover as to which group a new customer belongs to. Two approaches have been employed within machine learning to learn classifications. They are Neural Network based approaches and Induction based approaches. Both approaches have a number of advantages and disadvantages. Neural Networks may take longer to train than a rule induction approach but they are known to be better at learning to classify in situations where the data is noisy. However, as it is difficult to explain why a Neural Network made a particular classification they are dismissed as unsuitable for real Data Mining. Rule Induction based approaches to classification are normally Decision Tree based. Decision Trees can get very large and cumbersome when the training set is large, which is the case in Data Mining, and though they are not black boxes like Neural Networks become difficult to understand as well. This involves rules that associate one attribute of a relation to another. This involves rules that are based on temporal data. Suppose we have a database of natural disasters. From such a database if we conclude that whenever there was an earthquake in Los Angeles, the next day Mt. Kilimanjaro erupted, such a rule would be a sequence rule. Such rules are useful for making predictions which could be useful in making market gains or for taking preventive action against natural disasters. The factor that differentiates sequence rules from other rules is the temporal factor. This technique is based on two observations: A Data Mining Model 5. As mentioned in section 2, data stored in the real-world is full of anomalies that need to be dealt with before sensible discovery can be made. Alternatively, a Data Warehouse see section 8.

Chapter 2 : Chapter 2: Overview of the data mining process | Data Mining for Business Intelligence - Answ

Overview of the Data Mining Process In K aË™ re this chapter we give an overview of the steps involved in data mining, starting from a clear goal definition and ending with model deployment. The general steps are shown schematically in Figure

Data mining is used wherever there is digital data available today. Notable examples of data mining can be found throughout business, medicine, science, and surveillance. A common way for this to occur is through data aggregation. Data aggregation involves combining data together possibly from various sources in a way that facilitates analysis but that also might make identification of private, individual-level data deducible or otherwise apparent. Data may also be modified so as to become anonymous, so that individuals may not readily be identified. This indiscretion can cause financial, emotional, or bodily harm to the indicated individual. In one instance of privacy violation, the patrons of Walgreens filed a lawsuit against the company in for selling prescription information to data mining companies who in turn provided the data to pharmaceutical companies. Safe Harbor Principles currently effectively expose European users to privacy exploitation by U. The HIPAA requires individuals to give their "informed consent" regarding information they provide and its intended present and future uses. Use of data mining by the majority of businesses in the U. Copyright law[edit] Situation in Europe[edit] Due to a lack of flexibilities in European copyright and database law , the mining of in-copyright works such as web mining without the permission of the copyright owner is not legal. Where a database is pure data in Europe there is likely to be no copyright, but database rights may exist so data mining becomes subject to regulations by the Database Directive. On the recommendation of the Hargreaves review this led to the UK government to amend its copyright law in [36] to allow content mining as a limitation and exception. Only the second country in the world to do so after Japan, which introduced an exception in for data mining. However, due to the restriction of the Copyright Directive , the UK exception only allows content mining for non-commercial purposes. UK copyright law also does not allow this provision to be overridden by contractual terms and conditions. The European Commission facilitated stakeholder discussion on text and data mining in , under the title of Licences for Europe. As content mining is transformative, that is it does not supplant the original work, it is viewed as being lawful under fair use. Data mining and machine learning software. Public access to application source code is also available. Text and search results clustering framework. A chemical structure miner and web search engine. The Konstanz Information Miner, a user friendly and comprehensive data analytics framework. MEPX - cross platform tool for regression and classification problems based on a Genetic Programming variant. A software package that enables users to integrate with third-party machine-learning packages written in any programming language, execute classification analyses in parallel across multiple computing nodes, and produce HTML reports of classification results. A suite of libraries and programs for symbolic and statistical natural language processing NLP for the Python language. Open neural networks library. A component-based data mining and machine learning software suite written in the Python language. A programming language and software environment for statistical computing, data mining, and graphics. It is part of the GNU Project. An open source deep learning library for the Lua programming language and scientific computing framework with wide support for machine learning algorithms. A suite of machine learning software applications written in the Java programming language. Proprietary data-mining software and applications[edit] The following applications are available under proprietary licenses. OpenText Big Data Analytics: An environment for machine learning and data mining experiments. Visualisation-oriented data mining software, also for teaching. Marketplace surveys[edit] Several researchers and organizations have conducted reviews of data mining tools and surveys of data miners. These identify some of the strengths and weaknesses of the software packages. They also provide an overview of the behaviors, preferences and views of data miners. Some of these reports include: Report for Advanced Analytics as a market research assessment tool, it highlights both the diverse uses for advanced analytics technology and the vendors who make those applications possible. Recent-research Wiley Interdisciplinary Reviews:

Chapter 3 : What is the CRISP-DM methodology?

Data Overview of data mining process 2 Exercise You, as a company CEO, want to know the answers to the following questions. Which ones require a data.

Non-linear predictive models that learn through training and resemble biological neural networks in structure. Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution. Tree-shaped structures that represent sets of decisions. So these decisions generate rules for the classification of a dataset. A technique that classifies each record in a dataset based on a combination of the classes of the k record s most similar to it in a historical dataset. The extraction of useful if-then rules from data based on statistical significance. The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships. Systems that construct classifiers are one of the commonly used tools in data mining [8]. Such Systems take as input a collection of cases, each belonging to one of a small number of Classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs. The k-means algorithm is a simple iterative method to partition a given dataset into a user specified Number of clusters, k. Each data point is assigned to its closest centroid, with ties Broken arbitrarily. This results in a partitioning of the data. Each cluster representative is relocated to the center Mean of all data points assigned to it. The machine learning applications, Support Vector Machines SVM are considered A must tryâ€”it offers one of the most robust and accurate methods among all well-known Algorithms. Therefore it has a sound theoretical foundation, requires only a dozen examples for training, and is insensitive to the number of dimensions. One of the most popular data mining approaches is to find frequent item sets from a transaction Dataset and derive association rules. Therefore Finding frequent item sets is not trivial because of its combinatorial explosion. The EM algorithm; Finite mixture distributions provide a flexible and mathematical-based approach to the modeling and clustering of data observed on random phenomena. Therefore we focus here on the use of Normal mixture models, which can be used to cluster continuous data and to estimate the Underlying density function. The most popular Page Ranking algorithm issued by Google search engine. The algorithm assigns ranks for each hyperlink on the web. Based on this algorithm, they built the search engine Google, which has been a huge success. Nowadays every search engine has its own hyperlink based ranking method. Ensemble learning deals with methods which employ multiple learners to solve a problem. This generalization ability of an ensemble is usually significantly better than that of a single learner, so ensemble methods are very attractive. One of the simplest and rather trivial classifiers is the Rote classifier, which memorizes the entire training data and performs classification only if the attributes of the test object match one of the training examples exactly. This review would help the researchers to focus on the various issues of data mining. An overview of knowledge discovery database and data mining techniques has provided an extensive study on data mining techniques. Data mining is useful for both public and private sectors for finding patterns, forecasting, discovering knowledge in different domains such as finance, marketing, banking, insurance, health care and retailing.

Chapter 4 : What Is Data Mining?

- 1. Develop an understanding of the purpose of the data mining project. - 2. Obtain the dataset to be used in the analysis. - 3. Explore, clean, and preprocess the data.

Focus on large data sets and databases Data mining can answer questions that cannot be addressed through simple query and reporting techniques. Automatic Discovery Data mining is accomplished by building models. A model uses an algorithm to act on a set of data. The notion of automatic discovery refers to the execution of data mining models. Data mining models can be used to mine the data on which they are built, but most types of models are generalizable to new data. The process of applying a model to new data is known as scoring. For example, a model might predict income based on education and other demographic factors. Predictions have an associated probability How likely is this prediction to be true? Prediction probabilities are also known as confidence How confident can I be of this prediction? Some forms of predictive data mining generate rules, which are conditions that imply a given outcome. Rules have an associated support What percentage of the population satisfies the rule? Grouping Other forms of data mining identify natural groupings in the data. For example, a model might identify the segment of the population that has an income within a specified range, that has a good driving record, and that leases a new car on a yearly basis. Actionable Information Data mining can derive actionable information from large volumes of data. For example, a town planner might use a model that predicts income based on demographics to develop a plan for low-income housing. A car leasing agency might a use model that identifies customer segments to design a promotion targeting high-value customers. A general introduction to algorithms is provided in "Data Mining Algorithms". Data Mining and Statistics There is a great deal of overlap between data mining and statistics. In fact most of the techniques used in data mining can be placed in a statistical framework. However, data mining techniques are not the same as traditional statistical techniques. Traditional statistical methods, in general, require a great deal of user interaction in order to validate the correctness of a model. As a result, statistical methods can be difficult to automate. Moreover, statistical methods typically do not scale well to very large data sets. Statistical methods rely on testing hypotheses or finding correlations based on smaller, representative samples of a larger population. Data mining methods are suitable for large data sets and can be more readily automated. In fact, data mining algorithms often require large data sets for the creation of quality models. OLAP and data mining are different but complementary activities. OLAP supports activities such as data summarization, cost allocation, time series analysis, and what-if analysis. However, most OLAP systems do not have inductive inference capabilities beyond the support for time-series forecast. Inductive inference, the process of reaching a general conclusion from specific examples, is a characteristic of data mining. Inductive inference is also known as computational learning. OLAP systems provide a multidimensional view of the data, including full support for hierarchies. This view of the data is a natural way to analyze businesses and organizations. Data mining, on the other hand, usually does not have a concept of dimensions and hierarchies. Data mining and OLAP can be integrated in a number of ways. For example, data mining can be used to select the dimensions for a cube, create new values for a dimension, or create new measures for a cube. OLAP can be used to analyze data mining results at different levels of granularity. Data Mining can help you construct more interesting and useful cubes. For example, the results of predictive data mining could be added as custom measures to a cube. Such measures might provide information such as "likely to default" or "likely to buy" for each customer. OLAP processing could then aggregate and summarize the probabilities. Data Mining and Data Warehousing Data can be mined whether it is stored in flat files, spreadsheets, database tables, or some other storage format. The important criteria for the data is not the storage format, but its applicability to the problem to be solved. Proper data cleansing and preparation are very important for data mining, and a data warehouse can facilitate these activities. However, a data warehouse will be of no use if it does not contain the data you need to solve your problem. Oracle Data Mining requires that the data be presented as a case table in single-record case format. All the data for each record case must be contained within a row. Most typically, the case table is a view that presents the data in the required format for mining.

Data mining is a powerful tool that can help you find patterns and relationships within your data. But data mining does not work by itself. It does not eliminate the need to know your business, to understand your data, or to understand analytical methods. Data mining discovers hidden information in your data, but it cannot tell you the value of the information to your organization. You might already be aware of important patterns as a result of working with your data over time. Data mining can confirm or qualify such empirical observations in addition to finding new patterns that may not be immediately discernible through simple observation. It is important to remember that the predictive relationships discovered through data mining are not necessarily causes of an action or behavior. You can use this information to help you develop a marketing strategy. However, you should not assume that the population identified through data mining will buy the product because they belong to this population. Asking the Right Questions Data mining does not automatically discover solutions without guidance. The patterns you find through data mining will be very different depending on how you formulate the problem. To obtain meaningful results, you must learn how to ask the right questions. For example, rather than trying to learn how to "improve the response to a direct mail solicitation," you might try to find the characteristics of people who have responded to your solicitations in the past. Understanding Your Data To ensure meaningful data mining results, you must understand your data. Data mining algorithms are often sensitive to specific characteristics of the data: Oracle Data Mining can automatically perform much of the data preparation required by the algorithm. But some of the data preparation is typically specific to the domain or the data mining problem. At any rate, you need to understand the data that was used to build the model in order to properly interpret the results when the model is applied. The process flow shows that a data mining project does not stop when a particular solution is deployed. The results of data mining trigger new business questions, which in turn can be used to develop more focused models. Once you have specified the project from a business perspective, you can formulate it as a data mining problem and develop a preliminary implementation plan. For example, your business problem might be: Before building the model, you must assemble the data that is likely to contain relationships between customers who have purchased the product and customers who have not purchased the product. Data Gathering and Preparation The data understanding phase involves data collection and exploration. As you take a closer look at the data, you can determine how well it addresses the business problem. You might decide to remove some of the data or add additional data. This is also the time to identify data quality problems and to scan for patterns in the data. The data preparation phase covers all the tasks involved in creating the case table you will use to build the model. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, case, and attribute selection as well as data cleansing and transformation. Additionally you might add new computed attributes in an effort to tease information closer to the surface of the data. For example, rather than using the purchase amount, you might create a new attribute: Thoughtful data preparation can significantly improve the information that can be discovered through data mining. Model Building and Evaluation In this phase, you select and apply various modeling techniques and calibrate the parameters to optimal values. If the algorithm requires data transformations, you will need to step back to the previous phase to implement them unless you are using Oracle Automatic Data Preparation, as described in Chapter In preliminary model building, it often makes sense to work with a reduced set of data fewer rows in the case table , since the final case table might contain thousands or millions of cases. At this stage of the project, it is time to evaluate how well the model satisfies the originally-stated business goal phase 1. If the model is supposed to predict customers who are likely to purchase a product, does it sufficiently differentiate between the two classes? Is there sufficient lift? Are the trade-offs shown in the confusion matrix acceptable? Would the model be improved by adding text data? Should transactional data such as purchases market-basket data be included? Should costs associated with false positives or false negatives be incorporated into the model? See Chapter 5 for information about classification test metrics and costs. See Chapter 8 for information about transactional data. Knowledge Deployment Knowledge deployment is the use of data mining within a target environment. In the deployment phase, insight and actionable information can be derived from data. Deployment can involve scoring the application of models to new data , the extraction of model details for example the rules of a decision tree , or the integration of data mining models within

applications, data warehouse infrastructure, or query and reporting tools. Because Oracle Data Mining builds and applies data mining models inside Oracle Database, the results are immediately available. BI reporting tools and dashboards can easily display the results of data mining. Additionally, Oracle Data Mining supports scoring in real time:

Chapter 5 : Testing and Validation (Data Mining) | Microsoft Docs

An Overview of the Data Mining Process Tuesday The process of data mining allows a company to extract valuable insights and actionable information from data which.

Siddharth Mehta Overview Data Mining can be applied for a variety of purposes. Before one starts considering data mining as a probable solution, one should clearly understand the typical applications of data mining as well as the approach to develop data mining models in an enterprise. Having understood the fundamental considerations behind data mining, one can validate whether data mining would be the right solution for the problem. Also, one can assess the time, effort, infrastructure, and other resources that would be required to develop data mining models. Below are some of the basic concepts related to data mining. Explanation Below are the typical applications of data mining. Applications of Data Mining Forecasting: Estimating sales, predicting server loads or server downtime Risk and probability: Choosing the best customers for targeted mailings, determining the probable break-even point for risk scenarios, assigning probabilities to diagnoses or other outcomes Recommendations: Determining which products are likely to be sold together, generating recommendations Finding sequences: Analyzing customer selections in a shopping cart, predicting next likely events Grouping: Separating customers or events into cluster of related items, analyzing and predicting affinities The Data Mining development methodology is composed of different phases and each phase has its own considerations and purpose as stated below. Problem Definition What are you looking for? What types of relationships are you trying to find? Does the problem you are trying to solve reflect the policies or processes of the business? Do you want to make predictions from the data mining model, or are you just looking for interesting patterns and associations? Which outcome or attribute do you want to try to predict? What kind of data do you have and what kind of information is in each column? If there are multiple tables, how are the tables related? Do you need to perform any cleansing, aggregation, or processing to make the data usable? How is the data distributed? Is the data seasonal? Does the data accurately represent the processes of the business? Data Preparation Data can be scattered across a company and stored in different formats, or may contain inconsistencies such as incorrect or missing entries. The data might show that a customer bought a product before the product was offered on the market, or that the customer shops regularly at a store located 2, miles from her home. Data cleaning is not just about removing bad data or interpolating missing values, but about finding hidden correlations in the data, identifying sources of data that are the most accurate, and determining which columns are the most appropriate for use in analysis. Tools for data preparation: Exploration techniques include calculating the minimum and maximum values, calculating mean and standard deviations, and looking at the distribution of the data. You might determine by reviewing the maximum, minimum, and mean values that the data is not representative of your customers or business processes, and that you therefore must obtain more balanced data or review the assumptions that are the basis for your expectations. Standard deviations and other distribution values can provide useful information about the stability and accuracy of the results. A large standard deviation can indicate that adding more data might help you improve the model. Data that strongly deviates from a standard distribution might be skewed, or might represent an accurate picture of a real-life problem, but make it difficult to fit a model to the data. Data Mining Model Development You define the columns of data that you want to use by creating a mining structure. The mining structure is linked to the source of data, but does not actually contain any data until you process it. When you process the mining structure, Analysis Services generates aggregates and other statistical information that can be used for analysis. This information can be used by any mining model that is based on the structure. Before the structure and model is processed, a data mining model too is just a container that specifies the columns used for input, the attribute that you are predicting, and parameters that tell the algorithm how to process the data. Processing a model is often called training. Training refers to the process of applying a specific mathematical algorithm to the data in the structure in order to extract patterns. The patterns that you find in the training process depend on the selection of training data, the algorithm you chose, and how you have configured the algorithm. SQL Server contains many different algorithms, each suited to a different

type of task, and each creating a different type of model. Data Mining Model Validation Analysis Services provides tools that help you separate your data into training and testing datasets so that you can accurately assess the performance of all models on the same data. You use the training dataset to build the model, and the testing dataset to test the accuracy of the model by creating prediction queries. You can also test how well the models create predictions by using tools in the designer such as the lift chart and classification matrix. To verify whether the model is specific to your data, or may be used to make inferences on the general population, you can use the statistical technique called cross-validation to automatically create subsets of the data and test the model against each subset. Deploying and updating models Use the models to create predictions, which you can then use to make business decisions. Create content queries to retrieve statistics, rules, or formulas from the model. Use Integration Services to create a package in which a mining model is used to intelligently separate incoming data into multiple tables. Create a report that lets users directly query against an existing mining model. We will be focusing on the data mining model development, validation and deployment phase in the upcoming chapters assuming that we have the required data in place. Additional Information Consider reading this article to install the Adventure Works databases. It is recommended to install the version of the Adventure Works Data warehouse, though any version will work.

Chapter 6 : Data Mining - Microsoft Research

results of the data mining process, ensure that useful knowledge is derived from the data. Data mining is an extension of traditional data analysis and statistical approaches in that it incorporates analytical techniques drawn from a range of disciplines including, but not limited to.

The goals of this research project include development of efficient computational approaches to data modeling finding patterns , data cleaning, and data reduction of high-dimensional large databases. This enables database developers to easily access and successfully apply data mining technology in their applications. Current Status This is a long-term project. In the short term, the focus will be on automating the data mining process over data warehouses. This includes work in the following areas: Integration of data mining with database systems: Success of data mining as an enterprise technology crucially depends on seamless integration of this technology with enterprise databases. In this project, in collaboration with the SQL Server Product Group, we identify opportunities for new abstractions and interfaces that enable integration of data mining. Our future work will focus on exploiting data mining for advanced data summarization and also enable tighter coupling between database querying and data mining. Scalable Data Mining Algorithms: We are exploring scalable algorithms for modeling large databases. Specifically, we have focused on scalable decision tree algorithms for prediction, scalable probabilistic clustering algorithms, similarity detection algorithms between data objects, and mining sequence data. We are particularly interested in efficiently building data mining models in linear or near-linear time. The discovery part of the process â€" the part that finds gold among the gigabytes-is data mining. But before you can pull out your tin pan and shake it for gold, you need to gather your data into a data warehouse. Most major organizations have datawarehouses containing information about their clients, competitors and products. How will this help me? However, data mining might determine that the customers who spent the most dollars at your store bought the lowest selling items. Credit card companies have discovered another use for data mining. Before data mining, if you wanted to determine fraudulent transactions using a database you would query the database for all the transactions that had been determined fraudulent. Your next task would be to stare at hundreds of thousands of variables and try to decide which of the variables predicted fraud. Data mining algorithms structure the data and determine which attributes are relevant in a matter of minutes. SQL Server gets more power Until now, you had two choices: In the past, data mining tools used different data formats from those available in relational or OLAP multidimensional database systems. The data mining extensions in SQL Server will provide a common format for applications such as statistical analysis, pattern recognition, data prediction and segmentation methods, and visualization products. The data mining engine in SQL Server is a powerful platform. Though it ships with two algorithms, it is extensible and supports data mining algorithms that you might build. The two algorithms shipped with SQL Server are a scalable decision tree algorithm and a scalable clustering algorithm. A decision tree algorithm is meant to solve prediction problems. For instance, you might want to predict whether a high school student is going to go to college. If you have a database that contains information about people who did and did not go to college, the decision tree algorithm can use this data to learn rules to make predictions about new input. The rules can also tell you the percentage of probability of the prediction occurring. Someone using the system will be able to tell what rules where used to determine the prediction. Other predictive modeling methods, such as neural networks, are a bit more like a crystal ball, you just feed in the data and the prediction magically appears. Members of Microsoft Research and members of the SQL Server team at Microsoft came up with some clever techniques to pull the data out of SQL Server and quickly build decision trees from large sets of data. The clustering algorithms identify maximum similarities within a group, as well as maximum differences between groups. Customers may be grouped or segmented into those most likely to buy a certain product at certain times and under certain conditions. The resulting grouped clients are called clusters. Online stores who cluster their clients will recommend products to their customers based on past purchases. Statisticians have known about clustering algorithms for decades, however, most of the popular algorithms that are easy to implement will run quickly over small sets of data, but break down when applied to large sets. The main

problem is in the design. The algorithms run over and over until the groupings are found, and they may require many scans of the database at each iteration. If a business has a large database or one that is spread out over different servers, pulling the data together for even one customer is non-trivial. It could take days to obtain information about your market clusters. The scalable clustering algorithm in SQL Server clusters the database with one scan of the data. This helps address the computational difficulties of collecting data spread throughout an organization on different servers, since the data needs to be read only once. After that, the data mining specification does it for you. The algorithms also solve problems with high dimensional, sparse data. High dimensional data contains millions of data points in thousands of dimensions. Sparse data means that each entity has only a few of the characteristics that are being measured. Clustering documents is one application of this algorithm. This information will be useful to the thousands of dot coms hoping to get your business by serving up the content that you want when you need it, instead of making you slog through pages and pages of ever increasing data.

Chapter 7 : Data Mining Processes

The training set is used to train the data on various models. The validation set will test the trained data to see which model predicts the best. The test data gives an indication of how the model will perform with unknown examples.

Introduction to Data Mining Processes Data mining is a promising and relatively new technology. Data mining is defined as a process of discovering hidden valuable knowledge by analyzing large amounts of data, which is stored in databases or data warehouse , using various data mining techniques such as machine learning, artificial intelligence AI and statistical. Therefore, the needs for a standard data mining process increased dramatically. A data mining process must be reliable and it must be repeatable by business people with little or no knowledge of data mining background. As the result, in , a cross-industry standard process for data mining CRISP-DM first published after going through a lot of workshops, and contributions from over organizations. Next, we have to assess the current situation by finding the resources, assumptions, constraints and other important factors which should be considered. Then, from the business objectives and current situations, we need to create data mining goals to achieve the business objectives within the current situation. Finally, a good data mining plan has to be established to achieve both business and data mining goals. The plan should be as detailed as possible. Some important activities must be performed including data load and data integration in order to make the data collection successfully. Then, the data needs to be explored by tackling the data mining questions, which can be addressed using querying, reporting, and visualization. The outcome of the data preparation phase is the final data set. The data exploration task at a greater depth may be carried during this phase to notice the patterns based on business understanding. Modeling First, modeling techniques have to be selected to be used for the prepared dataset. Next, the test scenario must be generated to validate the quality and validity of the model. Then, one or more models are created by running the modeling tool on the prepared dataset. Finally, models need to be assessed carefully involving stakeholders to make sure that created models are met business initiatives. Evaluation In the evaluation phase, the model results must be evaluated in the context of business objectives in the first phase. In this phase, new business requirements may be raised due to the new patterns that have been discovered in the model results or from other factors. Gaining business understanding is an iterative process in data mining. The go or no-go decision must be made in this step to move to the deployment phase. Deployment The knowledge or information, which we gain through data mining process, needs to be presented in such a way that stakeholders can use it when they want it. Based on the business requirements, the deployment phase could be as simple as creating a report or as complex as a repeatable data mining process across the organization. From the project point of view, the final report of the project needs to summary the project experiences and review the project to see what need to improved created learned lessons.

Chapter 8 : KDD Process/Overview

Data Mining Tools: The Data Mining tools consist of the algorithms that automatically discover patterns from the pre-processed data. The tool chosen depends on the mining task at hand. The tool chosen depends on the mining task at hand.

It is important that you validate your mining models by understanding their quality and characteristics before you deploy them into a production environment. This section introduces some basic concepts related to model quality, and describes the strategies for model validation that are provided in Microsoft Analysis Services. For an overview of how model validation fits into the larger data mining process, see Data Mining Solutions. Methods for Testing and Validation of Data Mining Models There are many approaches for assessing the quality and characteristics of a data mining model. Use various measures of statistical validity to determine whether there are problems in the data or in the model. Separate the data into training and testing sets to test the accuracy of predictions. Ask business experts to review the results of the data mining model to determine whether the discovered patterns have meaning in the targeted business scenario All of these methods are useful in data mining methodology and are used iteratively as you create, test, and refine models to answer a specific problem. No single comprehensive rule can tell you when a model is good enough, or when you have enough data. Definition of Criteria for Validating Data Mining Models Measures of data mining generally fall into the categories of accuracy, reliability, and usefulness. Accuracy is a measure of how well the model correlates an outcome with the attributes in the data that has been provided. There are various measures of accuracy, but all measures of accuracy are dependent on the data that is used. In reality, values might be missing or approximate, or the data might have been changed by multiple processes. Particularly in the phase of exploration and development, you might decide to accept a certain amount of error in the data, especially if the data is fairly uniform in its characteristics. For example, a model that predicts sales for a particular store based on past sales can be strongly correlated and very accurate, even if that store consistently used the wrong accounting method. Therefore, measurements of accuracy must be balanced by assessments of reliability. Reliability assesses the way that a data mining model performs on different data sets. A data mining model is reliable if it generates the same type of predictions or finds the same general kinds of patterns regardless of the test data that is supplied. For example, the model that you generate for the store that used the wrong accounting method would not generalize well to other stores, and therefore would not be reliable. Usefulness includes various metrics that tell you whether the model provides useful information. For example, a data mining model that correlates store location with sales might be both accurate and reliable, but might not be useful, because you cannot generalize that result by adding more stores at the same location. Moreover, it does not answer the fundamental business question of why certain locations have more sales. You might also find that a model that appears successful in fact is meaningless, because it is based on cross-correlations in the data. Tools for Testing and Validation of Mining Models Analysis Services supports multiple approaches to validation of data mining solutions, supporting all phases of the data mining test methodology. Partitioning data into testing and training sets. Filtering models to train and test different combinations of the same source data. Measuring lift and gain. A lift chart is a method of visualizing the improvement that you get from using a data mining model, when you compare it to random guessing. Performing cross-validation of data sets Generating classification matrices. These charts sort good and bad guesses into a table so that you can quickly and easily gauge how accurately the model predicts the target value. Creating scatter plots to assess the fit of a regression formula. Creating profit charts that associate financial gain or costs with the use of a mining model, so that you can assess the value of the recommendations. These metrics do not aim to answer the question of whether the data mining model answers your business question; rather, these metrics provide objective measurements that you can use to assess the reliability of your data for predictive analytics, and to guide your decision of whether to use a particular iterate on the development process. The topics in this section provide an overview of each method and walk you through the process of measuring the accuracy of models that you build using SQL Server Data Mining.

Chapter 9 : Cross-industry standard process for data mining - Wikipedia

1. have data available in which the value of the outcome of interest is known which is called training data 2. training data are the data from which the classification or prediction algorithm "learns," or is "trained," about the relationship.

Your organisation may have competing objectives and constraints that must be properly balanced. What are the desired outputs of the project? For example, your primary goal might be to keep current customers by predicting when they are prone to move to a competitor. The plan should specify the steps to be performed during the rest of the project, including the initial selection of tools and techniques. Make sure that you are allowed to use the data. List the assumptions made by the project. These may be assumptions about the data that can be verified during data mining, but may also include non-verifiable assumptions about the business related to the project. It is particularly important to list the latter if they will affect the validity of the results. List the constraints on the project. These may be constraints on the availability of resources, but may also include technological constraints such as the size of data set that it is practical to use for modelling. List the corresponding contingency plans - what action will you take if these risks or events take place? This will generally have two components: A glossary of relevant business terminology, which forms part of the business understanding available to the project. A glossary of data mining terminology, illustrated with examples relevant to the business problem in question. This comparison should be as specific as possible. For example, you should use financial measures in a commercial situation. Determine data mining goals A business goal states objectives in business terminology. A data mining goal states project objectives in technical terms. Produce project plan Describe the intended plan for achieving the data mining goals and thereby achieving the business goals. Your plan should specify the steps to be performed during the rest of the project, including the initial selection of tools and techniques. Project plan - List the stages to be executed in the project, together with their duration, resources required, inputs, outputs, and dependencies. Where possible, try and make explicit the large-scale iterations in the data mining process, for example, repetitions of the modelling and evaluation phases. As part of the project plan, it is also important to analyze dependencies between time schedule and risks. Mark results of these analyses explicitly in the project plan, ideally with actions and recommendations if the risks are manifested. Decide at this point which evaluation strategy will be used in the evaluation phase. Your project plan will be a dynamic document. Specific review points for these updates should be part of the project plan. Here, for example, you select a data mining tool that supports various methods for different stages of the process. It is important to assess tools and techniques early in the process since the selection of tools and techniques may influence the entire project. This initial collection includes data loading, if this is necessary for data understanding. For example, if you use a specific tool for data understanding, it makes perfect sense to load your data into this tool. Record problems you encountered and any resolutions achieved. This will help both with future replication of this project and with the execution of similar future projects. Evaluate whether the data acquired satisfies your requirements. Distribution of key attributes for example, the target attribute of a prediction task Relationships between pairs or small numbers of attributes Results of simple aggregations Properties of significant sub-populations Simple statistical analyses These analyses may directly address your data mining goals. They may also contribute to or refine the data description and quality reports, and feed into the transformation and other data preparation steps needed for further analysis. Verify data quality Examine the quality of the data, addressing questions such as: Is the data complete does it cover all the cases required? Is it correct, or does it contain errors and, if there are errors, how common are they? Are there missing values in the data? If so, how are they represented, where do they occur, and how common are they? If quality problems exist, suggest possible solutions. Solutions to data quality problems generally depend heavily on both data and business knowledge. The criteria you might use to make this decision include the relevance of the data to your data mining goals, the quality of the data, and also technical constraints such as limits on data volume or data types. Note that data selection covers selection of attributes columns as well as selection of records rows in a table. This may involve selecting clean subsets of the data, the insertion of suitable defaults, or more ambitious techniques such as the estimation of missing data

by modelling. Consider any transformations of the data made for cleaning purposes and their possible impact on the analysis results. Construct required data This task includes constructive data preparation operations such as the production of derived attributes or entire new records, or transformed values for existing attributes. For example you might need to create records for customers who made no purchase during the past year. There was no reason to have such records in the raw data, but for modelling purposes it might make sense to explicitly represent the fact that particular customers made zero purchases. Integrate data These are methods whereby information is combined from multiple databases, tables or records to create new records or values. Each of these tables contains one record for each store. These tables can be merged together into a new table with one record for each store, combining fields from the source tables. If multiple techniques are applied, perform this task separately for each technique. Modelling technique - Document the actual modelling technique that is to be used. Record any assumptions made. For example, in supervised data mining tasks such as classification, it is common to use error rates as quality measures for data mining models. Therefore, you typically separate the dataset into train and test sets, build the model on the train set, and estimate its quality on the separate test set. Test design - Describe the intended plan for training, testing, and evaluating the models. Build model Run the modelling tool on the prepared dataset to create one or more models. Parameter settings - With any modelling tool there are often a large number of parameters that can be adjusted. List the parameters and their chosen values, along with the rationale for the choice of parameter settings. Models - These are the actual models produced by the modelling tool, not a report on the models. Model descriptions - Describe the resulting models, report on the interpretation of the models and document any difficulties encountered with their meanings. This task only considers models, whereas the evaluation phase also takes into account all other results that were produced in the course of the project. You should take the business objectives and business success criteria into account as far as you can here. Model assessment - Summarise the results of this task, list the qualities of your generated models e. Revised parameter settings - According to the model assessment, revise parameter settings and tune them for the next modelling run. Iterate model building and assessment until you strongly believe that you have found the best model s. Document all such revisions and assessments. Stage five â€" evaluation Evaluate your results Previous evaluation steps dealt with factors such as the accuracy and generality of the model. Another option is to test the model s on test applications in the real application, if time and budget constraints permit. Data mining results involve models that are necessarily related to the original business objectives and all other findings that are not necessarily related to the original business objectives, but might also unveil additional challenges, information, or hints for future directions. Assessment of data mining results - Summarise assessment results in terms of business success criteria, including a final statement regarding whether the project already meets the initial business objectives. Approved models - After assessing models with respect to business success criteria, the generated models that meet the selected criteria become the approved models. Review process At this point, the resulting models appear to be satisfactory and to satisfy business needs. It is now appropriate for you to do a more thorough review of the data mining engagement in order to determine if there is any important factor or task that has somehow been overlooked. This review also covers quality assurance issuesâ€"for example: Did we use only the attributes that we are allowed to use and that are available for future analyses? Review of process - Summarise the process review and highlight activities that have been missed and those that should be repeated. Determine next steps Depending on the results of the assessment and the process review, you now decide how to proceed. Do you finish this project and move on to deployment, initiate further iterations, or set up new data mining projects? You should also take stock of your remaining resources and budget as this may influence your decisions. List of possible actions - List the potential further actions, along with the reasons for and against each option. Decision - Describe the decision as to how to proceed, along with the rationale. If a general procedure has been identified to create the relevant model s , this procedure is documented here for later deployment. This is where predictive analytics really helps to improve the operational side of your business. Plan monitoring and maintenance Monitoring and maintenance are important issues if the data mining result becomes part of the day-to-day business and its environment. The careful preparation of a maintenance strategy helps to avoid unnecessarily long periods of incorrect usage of data mining results. In

order to monitor the deployment of the data mining result s , the project needs a detailed monitoring process plan. This plan takes into account the specific type of deployment. Monitoring and maintenance plan - Summarise the monitoring and maintenance strategy, including the necessary steps and how to perform them. Produce final report At the end of the project you will write up a final report. Depending on the deployment plan, this report may be only a summary of the project and its experiences if they have not already been documented as an ongoing activity or it may be a final and comprehensive presentation of the data mining result s. Final report - This is the final written report of the data mining engagement. It includes all of the previous deliverables, summarising and organising the results. Final presentation - There will also often be a meeting at the conclusion of the project at which the results are presented to the customer. Review project Assess what went right and what went wrong, what was done well and what needs to be improved. For example, any pitfalls you encountered, misleading approaches, or hints for selecting the best suited data mining techniques in similar situations could be part of this documentation.