Kind of data in data mining

We at an age are often referred to as the information age. In this information age, because we believe this information leads to power and success, and thanks to sophisticated technologies such as computers, satellites, etc., we have collected formidous amounts of information. Initially, with the adventure of computers and means for mass digital storage, we started collecting and storing all kinds of data, relying on the power of computers to help sort through this filling of information. Unfortunately, these massive collections of data stored on disappeared structures very rapidly became overwhelming. This initial disruption leads to the creation of structured databases and database management systems (DBMS). The effective database management systems were very important well for management of a large copy of data and especially for efficient and efficient retrieval of particular information from a large collection whenever needed. The proliferation of database management systems has also contributed to recent massive gatherings of all kinds of information. Today, we have far more information than we can handle: from business transactions and scientific data, to satellite photos, text reports and military intelligence. Recovering information is simply no longer enough to make decisions. Confronted with large collections of data, we have now created new needs to help us make better managed choices. These needs are automatic summary of data, the extraction of the stored information, and the discovery of patterns of previous data throughout editing. We've collected a myriad of data, from simple numeric measurements and text documents, to more complex information such as spatial data, multimedia channels, and hypertext documents. Here is a list that is not exclusive to a variety of information collected in digital forms of databases and in flat files. Business transaction: Every transaction in the business industry is (often) memorized for perpetuity. These transactions are usually time related to and can inter-business deals such as purchases, trades, banks, stock, etc., or intra-business operations such as management of in-house wars and assets. Large department stores, for example, thanks to the rampant use of bar codes, store millions of daily transactions represent often therapists of data. Warehouse space is not the biggest issue, as the cost of hard disk is continuously dropping, but the effective usage of the data in a reasonable time frame for competitive decision-making is definitely the most important problem to solve for businesses that struggle to survive in a highly competitive world. Scientific data: Whether in a Swiss accelerator counts particles count, in the Canadian forest study of a grizzly beads carrying radio, on an iceberg southern iceberg gathering data on oceanic activities, or at American University investigating human psychology, our society is surprisingly the column amount of scientific data that needs to be analyzed. Unfortunately, we can capture and store more new data faster than we can analyze the old data already accumulated. Medical and personal data: From continuing government personnel and customer records, very large collections of information are continuously gathered about individuals and groups. Governments, companies and organizations such as hospitals, are storing significant amounts of personal data to help them manage human resources, better understand a market, or simply help customertel. Whatever the issues on privacy of this type of data are often revealed, this information is collected, used and even shared. When corrected with this other data can be light on customer behavior and the like. Surveillance video and photos: With the amazing fall in video camera price, the video cameras become ubiquitous. Video tapes from surveillance cameras are usually recycled and thus the content is lost. However, there is a trend today to store the tapes and even digits for future use and analysis. Satellite essence: There is a countless number of satellites around the globe: some are geo-stationary above a region, and some is orbiting around the Earth, but all are sent a non-stop stream of data to the surface. NASA, which controls a large number of satellites, receives more data every second than what all NASA researchers and engineers can deal with. Many satellite photos and data are made public as soon as they receive in their hope that other researchers can analyze them. Game: Our society will collect a formidable amount of data and statistics on games, players and athletes. From hockey notes, backpacks and car-racing laps, to swim times, push boxer push and seat positions, all the data is stored. Commentators and journalists are using this information to report, but coaches and athletes wish to exploit this data to improve performance and better understand opponents. Digital media: The proliferation of cheap scanner, desktop video camera and digital camera is one of the causes of the explosion of digital media reposority. In addition, many radio stations, TV channels and movie studios are digitally audio and video collections to improve the management of their multimedia assets. Associations like the NHL and the NBA have already begun converting their big-game game collections into digital forms. CAD and data engineering software: There are a multitude of hardware design assistance (CAD) systems for architects of design buildings or engineers to conserve component systems or circuits. These systems are generating a formical amount of data. Moreover, software engineering is a similarly considerable data source code, function library, objects, etc., which need powerful tools for management and maintenance. Virtual World: There are many applications to make utilization of three-dimensional virtual space. These spaces and the objects they have are described with special languages such as VRML. Ideally, these virtual spaces are described in a way that they can share objects and locations. There are a remarkable number of virtual reality objects and reputable spaces available. Management of these repositories as well as content-based research and retrieval from those repositories are still search problems, while the size of their collections will continue to grow. Report text and memos (e-mail messages): Most of the communications in and between companies or search organizations or even private persons, are based on reports and memories of frequent text forms exchanged by e-mail. These messages are regularly stored in digital forms for future use and references create digital libraries. The World Wide Repositories: Since the inception of the World Wide Internet in 1993, documents all kinds of formats, content and description have been collected and inter-connected with links that make it greater repository of data ever built. Despite its dynamic and destroyer nature, its heterogeneous features, and very often redundancy and inconsistency, the World Wide Internet is the most important data collection regularly used for reference because of the broad variety of topics covered and the infinite contributions of resources and publishers. Many believe that the World Wide Web is becoming the compilation of human knowledge. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and possibly interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. Data Mining, also popularly known as Discovery of Knowledge Database (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discoveries in databases (or KDD) are often treated as senonymous, data mining is actually part of the knowledge discovery process. The following figure (Figure 1.1) shows data mining as a step in an incidental knowledge article process. The Discovery of Knowledge Database process comprises of a few steps leading from previous data collection to some forms of new knowledge. The item process consists of the following steps: Data cleaning: also known as clean data, it is a phase in which noise data and data are removed from the collection. Data integration: In this step, multiple data sources, often heterogeneous, can be combined into a common source. Data selection: in this step, relevant to the analysis to decide on and retrieve from the data collection. Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into the appropriate form for mining procedures. Data mining: It is the critical step in which intelligent techniques are applied to potentially useful extract patterns. Model assessment: At this stage, strictly interesting patterns represent the knowledge identified based on the measures provided. Knowledge representation: The final phase in which knowledge of discovery is visually represented to the user. This step bulk uses visualization techniques to help users understand and interpret the min data results. It is common to combine some of these steps together. For example, can clean data and data integration can be done simultaneously as a pre-processing phase to generate a data warehouse. Data selection and data transformation also must combine the consolidation location of the data is the result of the selection, or, such as for the case of data storage, the selection is done on transformed data. The KDD is an item process. Once the discovery knowledge is presented to the user, the evaluation measures can be improved, the mining can be more refined, new data can be selected or more transformed, or new data sources can be integrated, in order to get different, more appropriate results. Data mining exits its name in the similarity between searching for highly valued information in a large database and stone mining for a highly valued ore container. Both means either sites in a large number of materials or engineering the probation material exactly pinpoint where the values reside. It is, however, a misnomer, since gold mining of stones is usually called gold mining and not rock mining, so by analogy, data mining should be called mining knowledge instead. But data mining has become the custom theme of accepting, and very rapidly a trend that even overshadowed more general terms such as knowledgeable discovery in databases (KDD) that describe a more comprehensive process. Other similar terms refer to the data mining dictionary: dredging data, extracting knowledge and discovery patterns. In principle, data mining is not specific to a media or data type. Data mining should be applicable to any type of information repository. However, algorithms and approaches may differ when applied to different types of data. Indeed, the challenges presented by different types of data vary significantly. Data mining was put into use and studied for databases, including relationshipal databases, object-relation databases and object-oriented databases, data storage, transactional databases, non-structured storage and semi-structured such as the World Wide Web, advanced databases such as spatial databases, multimedia databases, time-series and textureal databases, and even flat files. Here are some examples of more details: Flat files: Flat files are actually the most common data source for data mining algorithms, especially at the search level. Flat files are simple data files in binary text or binary format with a known structure of the min data mining algorithm to be applied. The data in these records can be transactions, time-series data, scientific measurements, etc. Database relationships: Briefly, a relationships database consists of a set of tables that have either values of entity attributes, or values of attributes from entity relationships. Table contains columns and rows, where columns represent attributes and rows represent tuples. A couple of relationship tables correspond to either an object or a relationship between objects and is identified by a set of attribute values that represent a unique key. In Figure 1.2 we introduce some Customer Relations, Items, and Borrow represent business activities in a NouVideoStore video store. These relationships are just a subsets of what could be a database for the video store and are provided as an example. The language that is most commonly used for relationship databases is SQL, allowing retrieval and manipulation of the data stored in the tables, as well as the calculation of hungry functions such as average, sum, min, max and count. For example, a SQL query to select the grouped videos by category should be: SELECT COUNT(*) FROM Items WHERE TYPE= Video GROUP By Category. Mining algorithms using relationship databases can be more versatile than mining mining algorithms specifically written for flat files, since they can take advantage of the relationship database structure. While data mining can benefit from SQL for data selection, transformation and consolidation, it goes beyond what SQL could provide, such as prediction, compare, detect deviation, etc. Data warehouse: A data warehouse as a store, a data warehouse is collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the same unified schema. A data warehouse provides options for analyzing data from different sources under the same roof. Let us assume that NouVideoStore becomes a franchise in North America. Many video stores that are part of Company NouVideoStore can have different databases and different structures. If the executive at the company wants to access data from all stores for strategic decision-making, future direction, marketing, etc., it would be more appropriate to store all the data at one site with a homogeneous structure that allows interactive analysis. In other words, data from the different stores should be loaded, cleaned, transformed and integrated together. To facilitate decision-making and multi-dimensional views, data storage is usually modeled by a multi-dimensional data structure. 1.3 Shows an example of a three-dimensional subset of a data cube structure used for NouVideoStore storage. The figure shows summary rentals grouped by movie categories, then a cross table of rentals summarized by movie categories and time (in quarter). The data cube provides rentals to resume along three dimensions: Category, Time, and City. A cube contains cells that store values in some grown measures (in this case rental count), and special cells that summation store together dimensions. Each dimension in the data cube has a hierarchy of values for a single attribute. Because of the structures, the pre-computer data are included with the hierarchy attribute values of the dimensions, the data cube well article for fast interactive query and analysis of data at different designal levels, known as On-Line Analytical Processing (OPLA). OLAP operations allow the browsing of data at different levels of abstractions, such as drill-down, scroll,up, slices, ten, etc. Figure 1.4 shows the exercise-down (on the time dimension) and roll-up (on the location dimensions) operations. Database transaction: A transaction database is a set of transactions that represent transactions, each with a time stamp, an identifier and a set of items. Associated with the transaction records might also be descriptive data for the items. For example, in the case of the video store, the rental table as shown in Figure 1.5, represents the transaction database. Each record is a rental contract with a customer identifier, a date, and the list of renting items (i.e. video tapes, games, VCR, etc.). Since database relationships are not allowed necessary tables (i.e. a set as attribute value), transactions are usually stored in flat records or stored in two normalized transaction tables, one for transactions and one for the transaction items. One typical data analysis mining on the data is the so-called market analysis or association regulation in which associations between items performed together or in the study sequence. Multimedia database: Multimedia database includes video, images, audio and text media. They can be stored on extended object-relationshipal or object-oriented databases, or simply on a file system. Multimedia is characterized by its high dimensions, which makes data mining even more difficult. Data mining from multimedia repositories may require computer vision, computer graphics, image interpretation, and natural language processing methodology. Spatial databases: Spatial databases are databases that are, in addition to usual data, geographic information stores such as maps, and global or regional positioning. These spatial databases introduce new challenges to mining algorithms. Time-Series databases: Time-series databases contain time-related data such as stock market data or logged activity. These databases usually have a flow of new data comes in, which sometimes causes the need for a real-time challenge analysis. Data mining in these databases often include the study of trends and correlation between the evolution of different variables, as well as the prediction of trends and movements of the variables in time. Figure 1.7 shows some examples of time-series data. World Wide Internet: The World Wide Web is the most heterogeneous and dynamic reputable available. A large number of authors and publishers are continuously contributing to its growth and metamorphesis, and a massive number of users have access to its resources every day. Data of the World Wide Web site is organized into inter-connected documents. These documents can be text, audio, video, previous data all editing, and even applications. Designally, the World Wide Web consists of three major components: the content of the internet, the consistipated available documents; the structure of the internet, which covers the links and the relationship between documents; and usship of the internet, describing how and when resources are accessed. A fourth dimension can be added that relates the dynamic nature or evolution of the documents. Data mining in the World Wide Web, or Internet mining, tries to address all these problems and is often divided into mining web content, web mining structures and internet mining usage. The types of templates that can be discovered depend on employee data mining jobs. By and large, there are two types of data mining jobs: data mining description that describes the general properties of the existing data, and data preaching mining that attempts to make prediction based on inference on the data available. The data mining functionalities with the variety of knowledge discovered are shortly introduced in this list: characterization: Data characterization is a summary of general characteristics of objects in a target class, and produces so-called feature policies. Data that is relevant to a user-specified class are normally retrieved by a database query and run in a summary module to extract the essence of the data at different levels of abstraction. For example, one may want to characterize the TevideoStore customers who regularly rent more than 30 movies a year. With concept hierarchy on the attributes that describe the target class, the attribute-oriented method-oriented method can be used, for example, to carry out summary data. Note that with a data cube that contains the summary of data, simple OLAP operations fit the purpose of data characterization. Discrimination: Data discrimination generated so called discriminant rules and is basically the comparison of the general characteristics of objects between two classes referred to as the target class and the contract class. For example, one may want to compare the general features of the customers who rent more than movies in the past year and people with the rental account below 5. The techniques used for data discrimination are very similar to the techniques used for data characterization and the exception that data discrimination results include comparative measures. Association analysis: Analysis Association is the discovery of the commonly called association rules. It studies frequency of items that are performed simultaneously in transactional databases, and based on a supported threshide, identifying the item frequently. Another door, trust, which is the conditional probability that an item appears in a transaction when another item appears, is used in the main association rules. Association analysis is commonly used for market basket analysis. For example, it might be useful for aVideoStore manager to know what movies are often renting together or if there is a relationship between renting a certain type of movies and buying Popcorn or pop. The discovery rules are in the form: P -> Q[s,c], where P and Q are congainctions of value-pair attributes, and s (for support) is the probability that P and Q appear together in a transaction and c (for trust) is the conditional probability that Q appears in a transaction when P is present. For example, the hypothetical association rule: RentType(X, game) and age (X, 13-19)->Purchasing(X,Pop)[s=2%, c=55%] would indicate that 2% of transactions are considered to aging customers between 13 and 19% who rent a game and buy a pop, and that there is a certainty of 55% that teens who rent a toy also buy pop. : Classification analysis is the organization of data in the given classes. Also known as supervised classification, the classification is used to assign class labels to order the objects in the data collection. Approach classification normally uses a training set where all objects are already associated with known class labels. The classification algorithm learns in the training set and builds a model. The model is used to classify new objects. For example, after starting a credit policy, their manager TeVideoStore might analyze customers' vis-to-screw credit scores, and label accordingly their customers who received credit with three possible safe labels, risks and many risks. The classification analysis would generate a model that could use either accept or reject credit demand in the future.Prediction: Prediction has attracted considerable attention given the potential implications of successful forecast within a business context. There are two major types of prediction: one can either try to predict some available data values or trends annatant, or predict a class label for some data. The letter is bound to classification. Once a classification model is built based on a training set, the class label of an object must be anticipated based on the attribute values of the object and the class attribute values. Prediction is however most commonly referred to the forecast of missing numeric values, or increase/decrease the trend of time-related data. The biggest idea is to use a large number of past values to consider probable future values. Clustering: Similar to classification, clustering is the organization of data in classes. However, unlike classification, in downsides, class labels are unknown and it is up to the editing algorithm to discover acceptable classes. Clustering is also called classification without supervisor, because the classification is not dictated by the given class label. There are many approaches to regulate all based on the principle of maximizing the same way between the objects in a same class (intra-class similarities) and minimizing the same way between the objects of different classes (similar to inter-class classes). Outlier Analysis: Outliers are data elements that cannot be grouped into a given class or group. Also known as exceptions or astonishments, are often very important to identify. While outliers may be regarded as noise and disposal of some applications, they can reveal valuable knowledge in other domains, and so they can be very significant and analysis are important. Evolution and Deviation Analysis: Evolution and deviation analysis regarding the study of weather-related data changes in time. Evolution analysis models trends evolution in data, which consensus to characterize, compare, classify or clustering to time related data. Deviation analysis, on the other hand, considers the difference between measured values and expected values, and attempts to find the cause of the deviation from the anticipated values. It is common that users don't have a clear idea of the type of model they can discover or need to discover in the data at hand. It is therefore important to have a versatile and inclusive data mining system that enables the discovery of different types of knowledge and at different levels of abstraction. This also makes the interactivity an important type of a data mining system. Data mining enables the discovery of potentially useful and unknown knowledge. Whether the discovery knowledge is new, useful or interesting, is very subject and depends on the application and the user. It's certain that data mining can generate, or discovery, a large number of patterns or policies. In some cases the number of rules can reach the millions. One can even think of a meta-min phase in me the data mining overssized results. To reduce the number of models or discovery policies that have a high probability to be non-interesting, one has to put a measure on the models. However, this escalates the issue to complete. The user would want to discover all rules or templates, but only interesting ones. 19 How interesting an discovery is, often called interesting, can be based on quantific purpose elements such as the validity of the models when tested on new data with some degree of certainty, or on some deputy topics such as understanding of the models, gifts of their models, or utilities. Discovering patterns can also be found interesting if they confirm or validate a hypothesis sought to confirm or unplawfully contradict a common belief. This brings the issue of describing interesting discoveries, such as discovery meta-rule guides that describe the forms of rules prior to the discovery process, and language refinement interesting that interactively query the results for interesting patterns after phase of discovery. Typically, interesting measurements are based on doorstep set by the user. These doorstep define the complete of the discovery templates. Identifying and measuring the interesting of models and rules to discover, or to be discovered, is essential for the assessment of knowledge of minors and the KDD process as a whole. While some concrete measures exist, assessing the interesting in discovering knowledge is still an important research issue. There are many data mining systems available or have been developed. Some are specialized systems dedicated to a given data source or are confirmed to limit data mining functionalities, other are more versatile and comprehensive. Mining systems can be categorized according to various criteria among other classifications are the following: classification according to the type of minor data source: this data classification category mining system according to what type of data is handled such as spatial data, multimedia data, time-series data, data, World Wide Web, etc. Classifications according to the data model mapped on: This classification categories data mining system is based on the data model involved such as relationshipal databases, object-oriented databases, data storage, transactional, etc. Classification according to the King of Knowledge Discovery: This classification category data mining system is based on the type of knowledge discovery or mining functionality data, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be complete systems offering several mining data functionalities simultaneously. Classification according to mining techniques used: Mining Systems and providing different techniques. This classification categories data mining according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database-oriented or data storage-oriented, etc. The classification can also be taken into account the degree of user interaction involved in the data mining process such as research-driven systems, interactive exploratory systems, or autonomous systems. A complete system would provide a wide variety of mining techniques to fit different situations and options, and offer different degrees of user interaction. Reliable mining data algorithms techniques that have sometimes existed for many years, but have only recently been implemented as reliable and vocal tools that time and yet greater classic statistical methods. While Data Mining is still in its childbirth, it is becoming a trend with ubiquitous. Before data mining develops into a conventional, matured and trusted discipline, many are still annatant issues to be addressed. Some of these issues are addressed below. Note that these issues are not exclusive and do not order in any way. Security and social issues: Security is an important issue with any data collection shared with/or is intended to be used for strategic decision making. Additionally, when data is collected for customer profiles, understanding user understanding, correlation to personal data and other information, etc., large amounts of sensitive and private information about individuals or companies are gathered and stored. This becomes controversial to give the confidential nature of some of this data and access to illegal potential to the information. Moreover, data mining may disclose new implied knowledge about individuals or groups that could be against privacy policies, especially if there are potential diseminations of information discovery. Another problem that occurs from this concern is the appropriate usage of data mining. Due to the value of data, databases of all content types are regularly sold, and because of the competitive advantage that can be obtained from implicit knowledge discovery, some important information might be overwhelming, while other information could be widely distributed and used without control. Problem Client User: The knowledge uncovered by data mining tools is useful as long as it's interesting, and above all understood by the user. Good data visualization facilitates the interpretation of data mining results, as well as helps users better understand their needs. Many data explorer analysis tasks are significantly facilitated by the ability to view data in an appropriate visual presentation. There are many visualization ideas and proposals for efficient graphic data presentations. However, there is still much research and accomplished in order to find good visualization tools for big data that could be used to display and manipulate knowledge mining. The major issues related to user interfaces and visualization are real-estate screens, rendering information, and interactions. Interactivity with the data and mining results is critical since it provides means for the user to focus and refine the mining tasks, as well as the picture different angles and at different conceptive levels. Mining methodology problems: The following issues regarding the mining approach are implemented and the limitations. Topics such as versatility in the mining approach, the diversity of available data, the dimensions of the domain, the major analysis (when known), the assessment of the discovery knowledge, the os of background knowledge and metadata, the control and the noise touch of data, etc. are all examples that can dictate min methodology choices. For example, it is often desirable to have different mining methods available since different approaches can be done differently depending on the data at hand. Moreover, different approaches can suit and solve the user's needs a different way. Most algorithms assume the data must be noise-free. This is of course a strong assumption. Most data contain exceptions, invalid or incomplete information, etc., that may be complicated, if not analysis, the analysis process and in many cases compromised the accuracy of the results. As a consequence, pretreatment data (clean data and transformation) become relevant. It is commonly seen as lost time, but data clearing, as time-consuming and frustrating as it can be, is one of the most important phases of the knowledge discovery process. Data mining techniques should be able to handle noise of data or incomplete information. More than the size of data, the size of the search space is even more decisive for mining technical data. The size of the search space often depends on the number of dimensions in the domain space. The search space is usually growing exponentially when the number of dimensions increases. This is known as the Curse of Dimension. This curse affects so seriously the performance of some data mining approaches that it is becoming one of the most urgent problems to solve. Performance issues: Many artificial intelligence and statistical methods exist for data analysis and interpretation. However, these methods were often not designed for the large data sets min data is dealing with today. Terabyte sizes are common. This increases the problems of scalability and efficiency of the mining methods when processing considerably large data. Algorithms with exponential and even medium-order polynomial complexity cannot be in convenient use for data mining. Linear algorithms are usually the normal. In the same term, samples can be used for mining instead of the whole data. However, concerns such as completeness and choice of samples may arise. Other topics in the issue of performance are incremental upgrade, with parallel programming. There is no doubt that parallelism can help solve the size problem if the data can be subdivised and the results can be merged later. Incremental upgrade is important for merging results in parallel mining, or upgrade data results when new data become available without having to re-analyze the complete data. Problem Data Source: There are many issues related to the data sources, some are practical such as the diversity of data types, while others are philosophical like the glut problem. We certainly have an excess of data since we already have more data than we can handle and we still collect data at an even higher rate. If the spread of database management systems helped increase the gathering of information, the advantage of data mining is certainly encouraging more data harvesting. The current practice is to collect data as much as possible now and process it, or try to process it, later. The concern is whether we're collecting the right data into the appropriate amount, whether we know what we want to do with it, and whether we distinguish what important data and what data is overlooked. Regarding the practical issues related to source data, there is the subject of eterogeneous databases and focuses on various complex data. We're storing different types of data in a variety of repositories. It is difficult to expect a data mining system effectively and efficiently achieve good mining results on all types of data and sources. Different types of data and sources may require distinct algorithms and methodology. Currently, there is a focus on relationship databases and data storage, but other approaches need pioneers for other specific complex data types. A versatile data mining tool, for all data types, may not be realistic. Moreover, the proliferation of heterogeneous data sources, at structural and semantic level, possesses important challenges not only within the database community, but also within the data community min. References M. Chen, J. Han, and P. Yu. Data mining: An BEVISION from a database perspective. IEE Trans. Knowledge and Data Engineering, 8:866-883, 1996. Uthurusamy, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Progress in Discovery Knowledge and Data Mining. AAAI / MIT Press, 1996. W. Frawley, G. Piatetsky-Shapiro and C. J. Matheus, Knowledge Discovery in Database: A Glimpse of G. Piatetsky-Shapiro et al. (eds.), Discovery of Knowledge in Database. AAAI / MIT Press, 1991. Ann and M. Kamber. Data mining: Concepts and Techniques. Morgan Kaufmann, 2000. T. Imielinski and H. Mannila. A database perspective on knowledge discovery. Communications ACM, 39:58-64, 1996. G. Piatetsky-Shapiro, S. Fayad, and P. Smyth. From data mining to discovery knowledge: An AVISION. In U.M. Fayyad, et al. (eds.), advanced in Discovery Knowledge and Data Mining, 1-35. AAAI / MIT Press, 1996. G. Piatetsky - Shapiro and W. Frawley. Knowledge of Databases. AAAI / MIT Press, 1991. 1991.