UNDERSTANDING DATA MINING & ITS APPLICATION TO INTELLIGENCE ENVIRONMENTS

Charu Sharma*

Kanika Aggarwal**

ABSTRACT

This paper focuses on the data mining and the current trends associated with it. It begins with the overview of Data mining system and clarifies how data mining and knowledge discovery in databases are related both to each other and to related fields, such as machine learning, statistics, and databases. Various steps are involved in Knowledge discovery in databases (KDD) which helps to convert raw data into knowledge. Data mining is just a step in KDD which is used to extract interesting patterns from data that are easy to perceive, interpret, and manipulate. Several major kinds of data mining methods, including generalization, characterization, classification, clustering, association, evolution, pattern matching, data visualization, and meta-rule guided mining will be reviewed. Techniques for mining knowledge in different kinds of databases, including relational, transaction, object-oriented, spatial, and active databases, as well as global information systems, will be discuss. The explosive growth of databases makes the scalability of data mining techniques increasingly important. DM algorithms have the ability to rapidly mine vast amount of data. It also defines various Data Mining tools that are used to analyze different kinds of data. This paper also defines problems associated with data mining and applications of data mining in different fields.

*RIMT, Mandi Gobindgarh, Punjab.

**SCDL, Pune, Maharashtra.

I. INTRODUCTION

Data mining is the process of posing queries and extracting previously unknown information in the form of patterns, trends and structures from large quantities of often-diverse data. It integrates various technologies including database management, machine learning, statistics, parallel processing and visualization. We now have several commercial products and research prototypes. The reason for this explosion is because the supporting technologies are becoming mature and we now have ways of collecting, storing and organizing data to facilitate effective mining. Data mining outcomes include forming clusters as well as making associations, classifications, and correlations. Various techniques such as neural networks, decision tree and rue-based algorithms are being applied to obtain the desired data mining outcomes. Many of these techniques operate on relational databases in which the data are organized into tables. Current trends in data mining include mining unstructured data such as text, voice, and video; mining data in disturbed and heterogeneous databases; and mining data from the World Wide Web to help electronic commerce sites. Whereas data mining has produced numerous benefits, it can also cause serious security problems. Because of these data mining tools, users now have ways of extracting unauthorized information and making all kinds of correlations, however unintended and undesirable they may be. Therefore, security and privacy aspects of data mining are receiving some consideration, a trend that is likely to increase with the improvements in the capabilities of data mining tools and techniques.

II. THE DATA MINING PROCESS

Several steps must be performed to achieve successful data mining as illustrated in Figure 1, the first of which is data preparation. Having good data is essential for good mining. Data must be scrubbed and cleaned, and in some cases transferred to a data warehouse. Many individuals in organizations often may not know where the data they need are stored. Therefore, before mining the data, various sources of data have to be identified and possibly stored in databases. These heterogeneous data sources may be integrated into a warehouse. Inconsistencies and

uncertainties have to be eliminated at all levels of integration, including semantic inconsistencies. Various data cleaning and warehousing tools are

being developed for this purpose.

When the data-preparation step is either completed, or proceeds to an acceptable level, the next step is to determine the desired outcomes of data mining, such as clustering,

associations, and classification. The desired outcome will determine the selection of the appropriate tools to carry out mining. Actually, this step can be performed in parallel with the data-cleansing step in most cases. The selection of tools is largely a trial-and-error process and engineers involved in data mining will improve with experience. This is one reason why data mining is said to be an art**.** Applying the data mining tools and getting some results is the less complex task. More challenging is to make sense out of the data. This is called extracting the "golden nuggets" or "finding the needles in the haystack." An important step is to carry out some actions based on the results to determine whether the mining effort was worthwhile.

DATA SOURCES

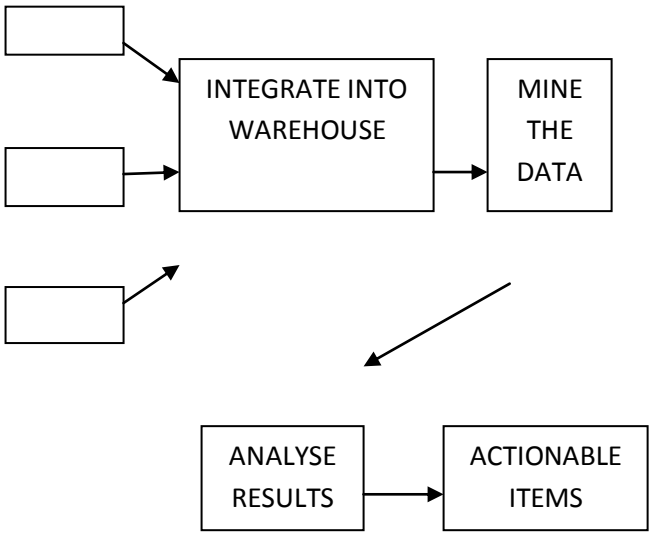


Figure 1. Process Of Data Mining

III. HOW IS DATA MINING USED

Data mining is used in a number of domains from financial, business, medical, advertising, to Command, Control, Communications, and Intelligence (C^3I). Many applications use data mining for marketing, promotions and sales, and some others use it for anomaly detection and making associations. To provide a better understanding of data mining, the following examples illustrate the practical uses of data mining:

- A credit bureau determines how its loans are processed by forming clusters of people with similar buying patterns and analyzing the clusters to see if a significant number of people in a cluster have defaulted on their previous payments. This information

may influence credit availability or interest rates for new borrowers whose characteristics match those in the high-risk clusters.

- A radiology department determines if an X-ray image is abnormal by training a neural network to learn normal images and flagging any image that deviates from the norm. This can help to decrease diagnostic errors of medical personnel who experience constant pressures in the face of a growing workload.
- An advertising agency sends customized promotional material in the mail by analyzing the sales of various department stores, forming clusters of people with similar buying patterns and sending appropriate promotional material to people depending on the clusters to which they belong. This can cut costs – Actionable Items for the agency by eliminating advertising with the lowest payoff.
- An intelligence agency makes associations between individuals by analyzing their travel schedules, spending patterns, and connections they make. Based on the mined information, the agency can target potential espionage or other illegal activity.

The last example shows how data mining may be applied in an intelligence environment. The following are more examples of data-mining applications for C^3I domains:

- A command and control research activity determines the best strategies for fighting a war by examining historical data, the current capabilities, and making predictions about the enemy resources and how they are likely to use them. A result of this analysis may be the following:
  - UK reconnaissance groups and US bomber groups work well together.
  - Operation RRR is not likely to succeed due to the strength and distribution of enemy resources.
  - The enemy commander is likely to order an air attack within the next 72 hours.
- A logistics agency determines the best strategy to build a new Air Force Base by analyzing historical data, current capabilities, and predicting future trends, not only in combat support and operations, but also by studying and predicting the world economy, weather, and population growth.
- A Federal agency applies data mining to determine intrusions in its computer and networked systems.
- An Intelligence agency determines secret collaborations of world leaders by analyzing various news stories.

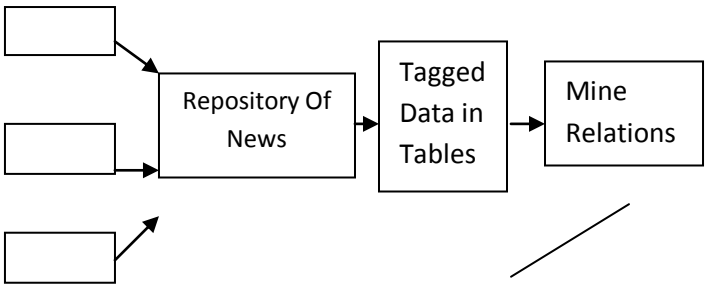
IV. THE DATA MINING SCENARIO FOR INTELLIGENCE APPLICATION

Now that we have presented an overview of what data mining is all about and how it may be used in a variety of domains including in C^3I, let us illustrate a sequence of steps for mining with a hypothetical scenario. Suppose an intelligence agency wants to analyze the various news stories. These news stories may be in the form of text, both from unclassified sources as well as from classified sources. These sources could be diverse, heterogeneous, and in different languages. Several options are available to tackle this problem. One is to mine the individual news stories from their original sources with a view toward fusing the mined information into a composite picture. However, with today's technology, we still have a long way to go with distributed data mining before it can be the technique of choice. Therefore, the preferred, current approach is to form a warehouse or some sort of repository of news stories and resolve the inconsistencies. This may involve translating foreign languages into a common language, such as English, either manually or with the help of a machine. Building a text warehouse is still in the early stages, and therefore, a considerable degree of manual intervention is needed.

Let us assume that we were successful in building a warehouse, which is essentially a repository for the news stories. The next step is to apply data mining to the news stories. This is essentially text mining. Therefore, various options are available. One option is to mine the text directly using text-mining tools and the other is to connect it into some structured form and mine the structured data. Since text-mining tools are in their infancy, with today's technology, one is very likely to choose the latter approach where text **is** converted to relational data, for example. This would mean tagging important entities in the text. Important entities would depend on the desired outcome. If we are mining to get information about associations between leaders, we need to extract the leaders, the places they travel, time of travel, as well as the reasons for travel, both stated and suspected. Various tagging tools are now available. So now we have the data in relational databases. The next step is to apply various mining tools. The challenges are to select the appropriate tool. If we are to find associations between individuals, then we need to select tools that form associations. Several commercial tools are now available for this purpose. Many of these tools output results in the form of "if-then" rules.

The mining tools may also output the accuracy of the rules as well as support for these rules. If the tool puts out hundreds of rules of the above form, it will be almost impossible for the

human to extract useful information. Therefore, techniques such as intelligence searching and pruning are used to determine which of the rules make sense based on the accuracy and support. Such searching and pruning techniques have come to be known as "market-basket analysis" techniques. These techniques were applied to analyse supermarket data to determine which items are purchased together. The same principle is applied here to determine the leaders who collaborate with each other. The goal is to find such leaders when such a pattern is not obvious to observers using unautomated inspection methods. For example, externally it may appear that Rita and John are combating while secretly they may be collaborating. This secret collaboration has to be uncovered by the data mining tool.



Collaborating Leaders

Person 1	Person 2	No.of times together
RITA	RICHA	10
JOHN	HARRY	25

Figure 2. Process Of Text Mining News Releases

When we get some meaningful results, we need to determine their validity. We have heard of plenty of false positives in data mining, and the last thing we want is for the Intelligence Agency to take some action based on wrong information. Current data mining tools are not sophisticated enough to produce the right information at the right time. Therefore, one suggestion is to implement a pilot project. For example, take news stories that are say two, three, and even five years old and carry out the

same procedure. Based on the rules obtained, specify some actions to be taken. But these actions, say, would be proposed for 1997 while we are in year 2000 now. So we analyze the proposed actions based on what we know today to see if this would be the right action. If so, we have some confidence in the results obtained with the tested technique as applied to today's news stories. Important data-mining applications are emerging for C^3I in the areas of link analysis and threat analysis. Link analysis is all about making associations and correlations between events and people, and following a chain of links so that one event leads to another. This type of analysis is fertile ground for the application of emerging research results in the field of conditional probability.

Threat analysis also would be extremely useful for information- warfare applications. Data mining also is under investigation for its utility to support counter terrorism activities. The Defense Research Projects Agency hosted a workshop on this topic in 1998 and is expected to start a research program in this area. For many of these applications, good text mining will be critical.

Another example of data mining is determining the changes that have occurred with, say, imagery data and determining whether the changes constitute a threat and whether further notification or action is warranted. Such mining is called "image mining." One approach here is to develop a neural network and train it to learn various types of imagery that we would consider normal. The neural network should also be able to detect normal changes such as seasonal changes to images. Therefore, when something unusual happens, such as "a ship is located" when it is not supposed to be in the picture, the neural network should flag this is an anomaly. The intelligence analyst will have to examine the change to determine if there is a reason to be alarmed. In this example, we have used neural network as a data mining technique, whereas in the news-story analysis example, we have used searching and pruning as data mining techniques.

V. FUTURE SCOPE

We can see that both text and image mining will be

important for C^3I applications. We can also expect relational data mining techniques to play a major role here. For example, much of the logistics data reside in tables; Therefore, mining relational databases will be necessary for such data. In many cases the data may be distributed across agencies and integrating them into a repository will be difficult due to organizational conflicts among other problems. Therefore, distributed data mining will be necessary. That is, we can apply data mining tools to the individuals data sources and then

merge the results to form the "big picture." As the volume of data-mining results grows, data mining experts may look to the field of data fusion for merging techniques.

Scalability is one of the major challenges in data mining and this will be the case for C^3I applications. Intelligence databases have grown to the petabyte size, and mining such large databases will be a challenge. That is, the data mining techniques have to handle large, heterogeneous and often multidimensional data sets.

Finally, we have to be concerned about security and privacy. If the data mining tools get into the wrong hands, we can expect disastrous situations. Security and privacy policies and techniques often conflict with policies for promoting data mining. Therefore, we need to balance these conflicting requirements and develop mining tools that are useful and yet do not violate security and privacy.

VI. REFERENCES

1. http://en.wikipedia.org/wiki/Data_mining
2. P Adriaan and D. Zaninge, Data Mining, Addison Wesley, 1996.
3. D. Bamber, I.R. Goodman and H. Nguyen, "Extension of the Concept of Propositional Deduction from Classical Logic to Probability: Part 1, an Overview of Probability-Selection Approaches," City, NJ, February 27- March 3,2000.131 M. Berry and G. Linoff, Dat.a Mining Techniques for Marketing, Sales and Customer Support, John Wiley and Sons, 1997.
4. M.G. Ceruti and M.N. Kamlzl, "Preprocessing and Integration of Data from Multiple Sources for Knowledge Discovery," International Journal on Artificial Intelligence Tools, (IJAIT), vol. 8, no. 2, pp. 152-177, June, 1999.
5. M.G. Ceruti and S.J. McCarthy, "Establishing a Data- Mining Environment for Wartime Event Prediction with an Object-Oriented Command and Control Database,"
6. C. Clifton and R. Steinheiser, "Data Mining on Text," Proceedings of the 22nd Annual IEEE International Conference.