CERIAS Tech Report 2001-93
Real-time data mining of multimedia objects
by Christopher Clifton
Center for Education and Research
Information Assurance and Security
Purdue University, West Lafayette, IN 47907-2086

Real-Time Data Mining of Multimedia Objects

**Bhavani Thuraisingham, Chris Clifton,
John Maurer**

*The MITRE Corporation
202 Burlington Road
Bedford, MA 01730-1420
(thura, clifton, johnm)@mitre.org*

Marion G. Ceruti
*Space and Naval Warfare Systems Center, San
Diego, D4121
53560 Hull Street, San Diego, CA, 92152-5001
ceruti@spawar.navy.mil*

Abstract

Whereas much of the previous work on data mining has focused on mining data in relational databases, in this paper we discuss mining objects. Object models are very popular for representing multimedia data and, therefore, we need to mine object databases to extract useful information from the large quantities of multimedia data. We first describe the motivation for multimedia data mining with examples and then discus object mining with focus on text, image, video and audio mining. We also address the need for real-time data mining for multimedia applications.

1. Introduction

Data mining is the process of posing queries and extracting information often previously unknown from large quantities of data. It integrates various technologies including database management, machine learning, statistics, parallel processing and visualization. During the past few years, data mining technology has exploded. We now have several commercial products and research prototypes. This explosion has occurred because the supporting technologies are becoming mature and we now have ways of collecting, storing and organizing the data so that they can be mined effectively. Data mining outcomes include forming clusters as well as making associations and correlations. Various techniques such as neural networks, Bayesian networks, decision tree and rule-based algorithms are being applied to obtain the desired data mining outcomes. Many of these techniques currently operate on relational databases where the data are organized in a set of tables. Current trends in data mining include mining unstructured data such as text, voice, and video, mining data in distributed and heterogeneous databases, and mining the World-Wide Web (WWW) data to help electronic commerce sites.

Engineers need to perform several steps for successful data mining The first is data preparation, as good data are essential for good mining. Data have to be scrubbed and cleaned, and in some cases, warehoused. Organizations often may not know where the data are stored. Therefore, before data mining is carried out, various sources of data have to be identified and possibly stored in databases. The heterogeneous data sources may be integrated into a warehouse. Inconsistencies and uncertainties must be eliminated. Various data cleaning and warehousing tools are being developed for this purpose. When the data preparation step is completed, the next step is to determine the desired outcomes, such as clustering, associations, and the subsequent selection of appropriate tools to carry out mining. This is largely a trial and error process. Engineers improve their performance with experience. Applying the data mining tools and getting some results is less complex task than making sense out of the data. This is called "extracting the golden nuggets" or "finding the needles in the haystack." The first step is to carry out some actions based on the results to determine whether the mining effort was worthwhile.

Whereas much of the previous work on data mining focused on mining relational databases, in this paper we discuss mining objects. Object models are very popular for multimedia data, and therefore, we need to mine object databases to extract useful information from the large quantities of multimedia data. The motivation for multimedia data mining is discussed in Section 2. Object modeling and mining for multimedia databases are the subjects of Section 3. We also discuss the issues of mining text, image, video and audio data in Section 3. Real-time and dependable computing aspects are described in Section 4. The paper is concluded in Section 5. For an overview of data mining we refer to [1], [2], and [3].

2. Motivation for Multimedia Data Mining

Data mining is used in a number of domains, from financial, medical, advertising, and Command, Control, Communications and Intelligence (C3I). Many applica-

360

tions use data mining for marketing, promotions and sales, and some others use it for detecting anomalies and making associations. To give the reader a better understanding of data mining, some examples are given below to illustrate the 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.
- 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.
- 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.
- An intelligence agency makes associations between individuals by analyzing their travel schedules, spending patterns, and connections they make.

Now that we have explained what data mining is and how it may be used in a variety of domains including in C3I, we now illustrate a sequence of steps to mining with a hypothetical scenario. Suppose an intelligence agency wants to analyze the various news stories as mentioned in the previous example. 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 ways are available to tackle this problem. One is to build a warehouse of news stories and resolve inconsistencies. Another is to mine the individual news stories and try and form a big picture. With today's technology, we still have a long way to go with distributed data mining. Therefore, the preferred current approach is to form a warehouse or some sort of repository of news stories. This may involve translating foreign languages into, say, English, either manually or with the help of a machine. Building a text warehouse is still in the early stages, and therefore, considerable manual intervention is needed.

Let us assume that we have successfully built 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 to us. One is to mine the text directly using text-mining tools and the other is to convert 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. Impor-

tant entities would depend on the object of our search. 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. 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 challenge now is 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 of the following form:

If John is present in Brussels, James will also be in Brussels at the same event.

If Jane is present in Boston, we can expect to see Mary the next day in London.

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 originally applied to analyze supermarket data to determine which items are purchased together. The same principle is applied here to determine which leaders collaborate with each other. The goal is to find such leaders when it is not obvious to the humans. For example, externally it may appear that John and Mary are combating while secretly they may be collaborating. This secret collaboration has to be uncovered by the data-mining tool. 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 an Intelligence Agency, for example, to take some action with grave consequences based on a false understanding of data mining results. Current data mining tools are not sophisticated enough to produce the right information at the right time, and to produce it in real time. Therefore, one approach is to try 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, whereas we are in year 2001 now. So analyze the proposed actions based on what we know today to see if this would be the right action. If so, we can have some confidence in the results obtained with today's news stories.

One of the important applications we see emerging with data mining for C3I applications is 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. Threat analysis would be extremely useful for information-warfare applications. Data mining also is being investigated to help with counter terrorism activities. The Defense Advanced Research Projects Agency (DARPA) hosted a workshop on this topic in 1998. The DARPA Rapid Knowledge Formation program has selected the domain of bio-terrorism for the subject matter of its knowledge bases (See, for example, [7]). For many of these applications, good text mining will be critical.

3. Multimedia Object Management and Mining

Data mining techniques typically operate on one of three models of data:

- Entity-based: Each data item is a distinct entity, with several attributes describing it. Data mining techniques look for patterns in the attributes that hold for many entities.
- Transaction-based: Data items are gathered into "transactions". Data mining techniques look for patterns among entities that occur in many transactions.
- Time-series data: A data item has a value at any given time. The data mining techniques look for patterns or repetitions in the values over time.

The type of data mining (e.g., clustering, classification, and association rules) determines what is done with the model. For example, classification in the entity-based model looks for patterns of attributes that distinguish "class members" from other entities.

To mine objects within the framework of existing data mining technology, we have to cast the objects into one of these models. For example, we may view a video object as a single entity; for data mining we need to extract a fixed set of "attributes" that describe that entity (length, number of scene changes, spectral characteristics of the frames, etc. We discuss some of the issues and possibilities below.

3.1 Object Modeling and Mining

We first discuss object modeling and mining and then examine text, image, audio and video mining. Both object and object-relational databases are suitable for managing multimedia data. We have had many debates in the past on which data model is suitable for multimedia databases and now we have more or less a consensus with both data models. Now, if multimedia data are to be mined, we need to mine object and object-relational databases. First we discuss modeling multimedia data with object and object-relational data models and then discuss mining aspects.

In representing multimedia data, several features have to be supported. First of all, we need a way to capture the complex data types and all the relationships between the data. Various temporal constructs such as play-before, play-after, play-together, etc. have to be captured. An appropriate data model is critical to represent a Multi-Media Database Management System (MM-DBMS). The relational, object-oriented, and object-relational data models have been examined to represent multimedia data. Some argue that relational models are better because they can capture relationships whereas others argue that object models are better because they represent complex structures. With an object-oriented data model, each object corresponds to an object in the object model. The attributes of an object may be represented as instance variables and will include time interval, frames, and content description. With the relational model, the object would correspond to an instance of a relation. However with atomic values, it will be difficult to capture the attributes of the instance. In the case of the object-relational model, the attribute value of an instance could be an object.

Languages such as Structured Query Language (SQL) are being extended for MM-DBMS. Others argue that object-oriented models are better as they can represent complex data types. It appears that both types of models have to be extended to capture the temporal constructs and other special features. Associated with a data model is a query language. The language should support the constructs needed to manipulate the multimedia database. For example, one may need to query to play frames 500 to 1000 of a video script.

In summary, several efforts are under way to develop appropriate data models for multimedia databases. Standards are also being developed. This is an area that will become mature within the next few years.

Now the challenge is to mine the data in the objects. One way is to extract relational data from multimedia data as we discussed in the case of text databases in Section 2. The other is to mine the object databases directly. Little research has been reported on mining object databases. First, we need to identify the types of mining that could be carried out. For example, we can find relationships between objects such as Object A and Object B by detecting associations. We can learn how to classify objects into predefined groups based on similarities in their properties. This is natural for objects as we can use the class concept to classify the objects. We can cluster objects. We can also do other types of mining such as estimation and prediction. The challenge is how to mine these objects.

We could develop techniques to retrieve values of the attributes and then find some relationships between the values. Here, we can use techniques for relational data mining. We can write additional methods to extract certain values and then do some analysis. We can also analyze the relationships between the object themselves. That is, we can use any of the techniques we have used before such as market basket analysis, neural networks, decision trees, and other methods for mining objects.

3.2 Text Mining

Much of the information is now in textual form. This could be data on the web or library data or electronic books, among others. One problem with text data is that text data are not structured as relational data. In many cases, text is unstructured and in some cases it is semistructured. An example of a semistructured dataset is an article that has a title, author, abstract, and paragraphs. The paragraphs are not structured, whereas the format is structured.

Information retrieval systems and text processing systems have been developed for more than a few decades. Some of these systems are quite sophisticated and can retrieve documents by specifying attributes or key words. Text-processing systems also can retrieve associations between documents. Frequently, we receive inquiries regarding the difference between information retrieval systems and text mining systems.

We define text mining as data mining on text data. Text mining involves extracting patterns and associations previously unknown from large text databases. The difference between text mining and information retrieval is analogous to the difference between data mining and database management. A distinct boundary does not exist. Some of the recent information retrieval and text-processing systems do discover associations between words and paragraphs, and therefore can be regarded as text mining systems.

Next, we examine the approaches to text mining. Note that many of the current tools and techniques for data mining work for relational databases. Even for data in object-oriented databases, rarely do we hear about data mining tools for such data. Therefore, current data mining tools cannot be applied directly to text data. Some of the current directions in mining unstructured data include the following examples.

- Extract data and/or metadata from the unstructured databases possibly by using tagging techniques, store the extracted data in structured databases, and apply data mining tools on the structured databases.
- Integrate data mining techniques with information retrieval tools so that appropriate data mining tools can be developed for unstructured databases.
- Develop data mining tools to operate directly on unstructured databases.

When converting text data into relational databases, one has to be careful not to lose key information. As we have stated above, without good data one cannot mine the data effectively and expect to get useful results. One needs to create a warehouse first before mining the converted database. This warehouse is essentially a relational database that has the essential data from the text data. In other words, one needs a transformer that takes a text corpus as input and outputs tables that have, for example, the keywords from the text.

As an example, in a text database that has several journal articles, one could create a warehouse with tables containing the following attributes: author, date, publisher, title, and keywords. From the keywords, one can form associations. The keywords in several articles could be "Belgium, nuclear weapons" and the keywords in other articles could be "Spain, nuclear weapons." The data miner could make the association that authors from Belgium and Spain write articles on nuclear weapons.

Note that we are only in the beginning of text mining. In the longer-term approach we would want to develop tools to mine text data directly. These tools have to read the text, understand the text, put out pertinent information about the text, and then make associations between different documents. We are far from developing such sophisticated text mining tools. However, the work in [4] is a step in the right direction. Some interesting early work on text mining was reported in [5].

3.3 Image Mining

If text mining is still in the early research stages, image mining is an even more immature technology. In this section, we will examine this area and discuss the current status and challenges.

Image processing has been around for some time. We have image processing applications in various domains including medical imaging for cancer detection, processing satellite images for space and intelligence applications, and also handling hyperspectral images. Images include maps, geological structures, biological structures, and many other entities. Image processing has dealt with areas such as detecting abnormal patterns that deviate from the norm, retrieving images by content, and pattern matching.

The main question here is what is image mining? How does it differ from image processing? Again we do not have clear-cut answers. One can say that while image processing is focusing on detecting abnormal patterns as well as retrieving images, image mining is all about finding unusual patterns. Therefore, one can say that image mining deals with making associations between different images from large image databases

Clifton et al. have begun work in image mining. Initially their plan was to extract metadata from images and then carry out mining on the metadata. This would essentially be mining the metadata in relational databases. However, after some consideration it was determined that a better approach was to mine the images directly. The challenge in this case was to determine the type of mining outcome that would be most suitable. One could mine for associations between images, cluster images, classify images, as well as detect unusual patterns. One area of research being pursued is to mine images to detect unusual features. So the approach is to develop templates that gen-

363

erate several rules about the images, and from there, apply the data mining tools to see if unusual patterns can be obtained. However, the mining tools will not tell us why these patterns are unusual. Note that detecting unusual patterns is not the only outcome of image mining; this is just the beginning. We need to conduct more research on image mining to determine whether data mining techniques could be used to classify, cluster, and associate images. Image mining is an area with applications in numerous domains including space, medicine, intelligence, and geoscience.

3.4 Video Mining

Mining video data is even more complicated than mining still image data. One can regard video as a collection of related still images, but is video something more? Video data management has been the subject of much research. The important areas include developing query and retrieval techniques for video databases, including video indexing, query languages, and optimization strategies. The first question one asks yet again is what is the difference between video information retrieval and video mining? Unlike image and text mining, we do not have any clear idea of what is meant by video mining. For example, one could examine video clips and find associations between different clips. Another example is to find unusual patterns in video clips. But how is this different from finding unusual patterns in images? The first step to successful video mining is to have a good handle on image mining.

We examine pattern matching in video databases. Should one have predefined images and then match these images with the video data? Can we perform pattern recognition in video data by specifying for the object of the search and extracting features from the video data? If this is video information retrieval what then is mining video data? To be consistent with our terminology we can say that finding correlations and patterns previously unknown from large video databases is video mining. By analyzing a video clip or multiple video clips, one comes to conclusions about some unusual behavior. People featured in the video who were not expected to be there, yet have occurred two or three times could mean something significant. Another way to look at the problem is to capture the text in video format and try and make the associations one would carry out with text but use the video data instead.

Unlike text and image mining where our ideas have been less vague, the discussion here on video mining is quite preliminary. This is mainly because so little known about video mining. Even the term "video mining" is something very new, and to date we do not have any concrete results reported on this. We do have a lot of information on analyzing video data and producing summaries. One could mine these summaries, which would amount to mining text A good example of this effort is the work by Merlino et al. on summarizing video news [6]. Converting

the video-mining problem to a text-mining problem is reasonably well understood. However, the challenge is mining video data directly, and more importantly, knowing what we want to mine. With the emergence of the web, video mining becomes even more important.

3.5 Audio Mining

Since audio is a continuous media type like video, the techniques for audio information processing and mining are similar to video information retrieval and mining. Audio data could be in the form of radio, speech, or spoken language. Even television news has audio data, and in this case audio data may have to be integrated with video and possibly text to capture the annotations and captions.

To mine audio data, one could convert the spoken word into text using speech transcription techniques and other techniques such as keyword extraction, and then mine the text data. On the other hand, audio data could also be mined directly by using audio information processing techniques and then mining selected audio data. In general, audio mining is even more primitive than video mining. Whereas a few papers have appeared on text mining and even fewer on image and video mining, work on audio mining is just beginning.

4. Real-Time and Dependable Computing for Multimedia Data Management and Mining

In real-time database system, the queries and transactions have to meet timing constraints. A strong connection has been observed between real-time database systems and multimedia database systems. For example, the audio and video data have to be synchronized. This means that certain timing constraints are imposed on the data and these timing constraints must be met. If not, one could hear Jane's voice with Robert speaking on the video script.

The issue is whether the timing constraints are hard, soft, or firm. If they are hard, then they have to be met. Otherwise, a catastrophic situation could develop. If they are firm, then they should to be met, but no serious consequence is associated with a missed deadline. If they are soft, meeting the constraints will enhance the overall result, but a missed deadline will not produce any unacceptable effects. After the type of constraints is determined for an application, techniques are needed to handle these constraints. Real-time scheduling techniques may help here.

Now, darts mining often is associated with non-real-time processing. For example much of the analysis of logistics data and other related data are carried out for non-real-time applications. The question is do we need to perform data mining for real-time applications? Although we are a long way from real-time data mining, the answer is yes. This is because often we need answers of analysis in real-time. If it is too late, then the answers may not be use-

ful. This is especially the case of financial and intelligence applications when timely decisions must be made to capture an advantage. For example, we need the analysis of financial data within a certain time to make effective decisions. In the same way, we need the answers to whether certain enemy threats are real in real-time. Otherwise, the results may be too late to prevent disaster.

How we do carry out real-time data mining? We need time constraints on the queries. Here we can use many of the techniques for processing real-time queries and transactions; we also need time constraints on the analysis. For example, when we are using market-basket analysis techniques, we need to assign time constraints on returning associations. This way we may not be able to obtain all associations, but some associations. That is, we need to incorporate quality of service primitives here. Do we need all answers in five hours or three answers in five minutes?

It is the same situation as one in which we would need dependable computing. However, the "flavor" is different here. In addition to traditional fault tolerant features, we need answers to questions such as:

- How accurate are the results?
- What is the confidence we have?
- If we have obtained incorrect answers, how do we correct the mistakes?

A considerable amount of research needs to be done here.

5. Summary and Trends

Whereas much of the previous work on data mining focuses on relational data mining, in this paper we discuss mining objects. Object models are very popular for representing multimedia data, and therefore, we need to mine object databases to extract useful information from the large quantities of multimedia data. We first described the motivation for multimedia-data mining with examples and then discussed object mining with focus on text, image, video, and audio mining. We also addressed need for real-time data mining for multimedia applications.

We have discussed some of the several areas that need work, which are summarized below.

- How do we mine objects? We gave some directions in this paper. How do we apply the data mining techniques such as market basket analysis for objects?
- How do we model text, image, audio, and video as objects and then mine the objects? This paper gave some preliminary directions.
- How can we incorporate timing constraints into data mining?
- How do we handle incorrect results? We discussed some dependability aspects here.

References

[1] Bhavani Thuraisingham, Data Mining: Technologies, Techniques, Tools and Trends, CRC Press, 1998

[2] M. Berry and G. Linoff, Data Mining Techniques for Marketing, Sales and Customer Support, John Wiley and Sons, 1997

[3] P Adriaan and D. Zaninge, Data Mining, Addison Wesley, 1996

[4] Dick Tsur, Jeffrey D. Ullman, Serge Abiteboul, Chris Clifton, Rajeev Motwani, Svetlozar Nestorov, and Arnon Rosenthal, Query Flocks: A Generalization of Association Rule Mining, in *Proceedings of the 1998 ACM SIGMOD Conference on Management of Data*, Seattle, WA, June 2-4, 1998

[5] R. Feldman and I. Dagan, Knowledge Discovery in Textual Databases (KDT), in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, AAAI, Montreal, Canada, August 1995.

[6] Mark Maybury, Andy Merlino, and Daryl Morey, Broadcast News Navigation using Story Segments, in Proceedings of the ACM International Multimedia Conference, Seattle, Washington, November 1997.

[7] M.G. Ceruti, C. Anken, A. D. Lin, and S.H. Rubin, Applications of High-performance Knowledge-Based Technology, In *Proceedings of the 2000 IEEE International Conference on Systems, Man and Cybernetics (SMC 2000)*, pp. 1965-1971, October 2000.