NEW DEVELOPMENTS IN DATA MINING: A REVIEW OF THE KEY DATA MINING TECHNOLOGIES AND APPLICATIONS FOR THE 21ST CENTURY

Jeffrey Hsu, Information Systems, Silberman College of Business Administration, Fairleigh Dickinson University, 285 Madison Avenue, Madison NJ 07940, jeff@fdu.edu

ABSTRACT

This paper discusses a number of technologies, approaches, and research areas which have been identified as having critical and future promise in the field of data mining. Among the areas which are being developed, investigated, and applications identified for include hypertext and hypermedia data mining, ubiquitous data mining (UDM), phenomenal data mining, distributed and collective data mining, time series, constraint-based, spatial/geographic, and related methods.

Keywords: data mining; distributed/collective, ubiquitous, hypertext/hypermedia, spatial/geographic, time series, constraint-based.

THE PRESENT AND THE FUTURE

The field of data mining and knowledge discovery in databases (KDD) has been growing in leaps and bounds, and has shown great potential for the future (5). The purpose of this chapter is to survey many of the future trends in the field of data mining, with a focus on those which are thought to have the most promise and applicability to future data mining applications. As mentioned previously, the field of data mining is very broad, and there are many methods and technologies which have become dominant in the field. Not only have there been developments in the "traditional" areas of data mining, there are other areas which have been identified as being especially important as future trends in the field. These are the subject of the upcoming sections and form the major emphasis of this paper.

DISTRIBUTED/COLLECTIVE DATA MINING

One area of data mining which is attracting a good amount of attention is that of distributed and collective data mining. Much of the data mining which is being done currently focuses on a database or data warehouse of information which is physically located in one place. However, the situation arises where information may be located in different places, in different physical locations. This is known generally as distributed data mining (DDM). Therefore, the goal is to effectively mine distributed data which is located in heterogeneous sites. Examples of this include biological information located in different databases, data which comes from the databases of two different firms, or analysis of data from different branches of a corporation, the combining of which would be an expensive and time-consuming process.

Distributed data mining (DDM) is used to offer a different approach to traditional approaches analysis, by using a combination of localized data analysis, together with a "global data model. " In more specific terms, this is specified as performing local data analysis for generating partial data models, and combining the local data models from different data sites in order to develop the global model. This global model combines the results of the separate analyses. Often the global model produced, especially if the data in different locations has different features or

characteristics, may become incorrect or ambiguous. This problem is especially critical when the data in distributed sites is heterogeneous rather than homogeneous. These heterogeneous data sets are known as vertically partitioned datasets. An approach proposed by Kargupta et al. (9) speaks of the collective data mining (CDM) approach, which provides a better approach to vertically partitioned datasets, using the notion of orthonormal basis functions, and computes the basis coefficients to generate the global model of the data (9).

UBIQUITOUS DATA MINING (UDM)

The advent of laptops, palmtops, cell phones, and wearable computers is making ubiquitous access to large quantity of data possible. Advanced analysis of data for extracting useful knowledge is the next natural step in the world of ubiquitous computing. Accessing and analyzing data from a ubiquitous computing device offer many challenges. For example, UDM introduces additional cost due to communication, computation, security, and other factors. So one of the objectives of UDM is to mine data while minimizing the cost of ubiquitous presence. Human-computer interaction is another challenging aspect of UDM. Visualizing patterns like classifiers, clusters, associations and others, in portable devices are usually difficult. The small display areas offer serious challenges to interactive data mining environments. Data management in a mobile environment is also a challenging issue. Moreover, the sociological and psychological aspects of the integration between data mining technology and our lifestyle are yet to be explored. The key issues to consider include theories of UDM, advanced algorithms for mobile and distributed applications, data management issues, mark-up languages, and other data representation techniques; integration with database applications for mobile environments, architectural issues: (architecture, control, security, and communication issues), specialized mobile devices for UDM, software agents and UDM (Agent based approaches in UDM, agent interaction---cooperation, collaboration, negotiation, organizational behavior), applications of UDM (Application in business, science, engineering, medicine, and other disciplines), location management issues in UDM and technology for web-based applications of UDM (10).

HYPERTEXT AND HYPERMEDIA DATA MINING

Hypertext and hypermedia data mining can be characterized as mining data which includes text, hyperlinks, text markups, and various other forms of hypermedia information. As such, it is closely related to both web mining, and multimedia mining, which are covered separately in this section, but in reality are quite close in terms of content and applications. While the World Wide Web is substantially composed of hypertext and hypermedia elements, there are other kinds of hypertext/hypermedia data sources which are not found on the web. Examples of these include the information found in online catalogues, digital libraries, online information databases, and the like. In addition to the traditional forms of hypertext and hypermedia, together with the associated hyperlink structures, there are also inter-document structures which exist on the web, such as the directories employed by such services as Yahoo! (www.yahoo.com) or the Open Directory project (http://dmoz.org) These taxonomies of topics and subtopics are linked together to form a large network or hierarchical tree of topics and associated links and pages.

Some of the important data mining techniques used for hypertext and hypermedia data mining include classification (supervised learning), clustering (unsupervised learning), semi-structured learning, and social network analysis. In the case of classification, or supervised learning, the process starts off by reviewing training data in which items are marked as being part of a certain

class or group. This data is the basis from which the algorithm is trained. One application of classification is in the area of web topic directories, which can group similar-sounding or spelled terms into appropriate categories, so that searches will not bring up inappropriate sites and pages. The use of classification can also result in searches which are not only based on keywords, but also on category and classification attributes. Methods used for classification include naive Bayes classification, parameter smoothing, dependence modeling, and maximum entropy (2). Unsupervised learning, or clustering, differs from classification in that classification involved the use of training data, clustering is concerned with the creation of hierarchies of documents based on similarity, and organize the documents based on that hierarchy. Intuitively, this would result in more similar documents being placed on the leaf levels of the hierarchy, with less similar sets of document areas being placed higher up, closer to the root of the tree. Techniques which have been used for unsupervised learning include k-means clustering, agglomerative clustering, random projections, and latent semantic indexing. Semi-supervised learning and social network analysis are other methods which are important to hypermedia-based data mining. Semi-supervised learning is the case where there are both labeled and unlabeled documents, and there is a need to learn from both types of documents. Social network analysis is also applicable because the web is considered a social network, which examines networks formed through collaborative association, whether it be between friends, academics doing research or service on committees, and between papers through references and citations. Graph distances and various aspects of connectivity come into play when working in the area of social networks (16). Other research conducted in the area of hypertext data mining include work on distributed hypertext resource discovery (3).

MULTIMEDIA DATA MINING

Multimedia Data Mining is the mining and analysis of various types of data, including images, video, audio, and animation. The idea of mining data which contains different kinds of information is the main objective of multimedia data mining (22). As multimedia data mining incorporates the areas of text mining, as well as hypertext/hypermedia mining, these fields are closely related. Much of the information describing these other areas also apply to multimedia data mining. This field is also rather new, but holds much promise for the future. Multimedia information, because its nature as a large collection of multimedia objects, must be represented differently from conventional forms of data. One approach is to create a multimedia data cube which can be used to convert multimedia-type data into a form which is suited to analysis using one of the main data mining techniques, but taking into account the unique characteristics of the data. This may include the use of measures and dimensions for texture, shape, color, and related attributes. In essence, it is possible to create a multidimensional spatial database. Among the types of analyses which can be conducted on multimedia databases include associations, clustering, classification, and similarity search. Another developing area in multimedia data mining is that of audio data mining (mining music). The idea is basically to use audio signals to indicate the patterns of data or to represent the features of data mining results. The basic advantage of audio data mining is that while using a technique such as visual data mining may disclose interesting patterns from observing graphical displays, it does require users to concentrate on watching patterns, which can become monotonous. But when representing it as a stream of audio, it is possible to transform patterns into sound and music and listen to pitches, rhythms, tune, and melody in order to identify anything interesting or unusual. Its is possible not only to summarize melodies, based on the approximate patterns that repeatedly occur in the

segment, but also to summarize style, based on tone, tempo, or the major musical instruments played (22, 5).

SPATIAL AND GEOGRAPHIC DATA MINING

The data types which come to mind when the term data mining is mentioned involves data as we know it—statistical, generally numerical data of varying kinds. However, it is also important to consider information which is of an entirely different kind—spatial and geographic data which could contain information about astronomical data, natural resources, or even orbiting satellites and spacecraft which transmit images of earth from out in space. Much of this data is image-oriented, and can represent a great deal of information if properly analyzed and mined . A definition of spatial data mining is as follows: "the extraction of implicit knowledge, spatial relationships, or other patterns not explicitly stored in spatial databases." Some of the components of spatial data which differentiate it from other kinds include distance and topological information, which can be indexed using multidimensional structures, and required special spatial data access methods, together with spatial knowledge representation and data access methods, along with the ability to handle geometric calculations.

Analyzing spatial and geographic data include such tasks as understanding and browsing spatial data, uncovering relationships between spatial data items (and also between non-spatial and spatial items), and also analysis using spatial databases and spatial knowledge bases. The applications of these would be useful in such fields as remote sensing, medical imaging, navigation, and related uses. Some of the techniques and data structures which are used when analyzing spatial and related types of data include the use of spatial warehouses, spatial data cubes and spatial OLAP. Spatial data warehouses can be defined as those which are subject-oriented, integrated, nonvolatile, and time-variant. Some of the challenges in constructing a spatial data warehouse include the difficulties of integration of data from heterogeneous sources, and also applying the use of on-line analytical processing which is not only relatively fast, but also offers some forms of flexibility. In general, spatial data cubes, which are components of spatial data warehouses, are designed with three types of dimensions and two types of measures. The three types of dimensions include the nonspatial dimension (data which is nonspatial in nature), the spatial to nonspatial dimension (primitive level is spatial but higher level-generalization is nonspatial), and the spatial-to-spatial dimension (both primitive and higher levels are all spatial). In terms of measures, there are both numerical (numbers only), and spatial (pointers to spatial object) measured used in spatial data cubes (18, 23). Aside from the implemention of data warehouses for spatial data, there is also the issue of analyses which can be done on the data. Some of the analyses which can be done include association analysis, clustering methods, and the mining of raster databases There have been a number of studies conducted on spatial data mining (1, 12, 20).

TIME SERIES/SEQUENCE-BASED DATA MINING

Another important area in data mining centers on the mining of time series and sequence-based data. Simply put, this involves the mining of a sequence of data, which can either be referenced by time (time-series, such as stock market and production process data), or is simply a sequence of data which is ordered in a sequence. In general, one aspect of mining time series data focuses on the goal of identifying movements or components which exist within the data (trend analysis). These can include long-term or trend movements, seasonal variations, cyclical variations, and random movements (5).

Other techniques which can be used on these kinds of data include similarity search, sequential pattern mining, and periodicity analysis. *Similarity search* is concerned with the identification of a pattern sequence which is close or similar to a given pattern, and this form of analysis can be broken down into two subtypes: whole sequence matching and subsequence matching. Whole sequence matching attempts to find all sequences which bear a likeness to each other, while subsequence matching attempts to find those patterns which are similar to a specified, given sequence. *Sequential pattern mining* has as its focus the identification of sequences which occur frequently in a time series or sequence of data. This is particularly useful in the analysis of customers, where certain buying patterns could be identified, such as what might be the likely follow-up purchase to purchasing a certain electronics item or computer, for example. *Periodicity analysis* attempts to analyze the data from the perspective of identifying patterns which repeat or recur in a time series. This form of data mining analysis can be categorized as being full periodic, partial periodic, or cyclic periodic. In general, full periodic is the situation where all of the data points in time contribute to the behavior of the series. This is in contrast to partial periodicity, where only certain points in time contribute to series behavior. Finally, cyclical periodicity relates to sets of events which occur periodically (5, 6, 8, 11).

CONSTRAINT-BASED DATA MINING

Many of the data mining techniques which currently exist are very useful but lack the benefit of any guidance or user control. One method of implementing some form of human involvement into data mining is in the form of constraint-based data mining. This form of data mining incorporates the use of constraints which guides the process. Frequently this is combined with the benefits of multidimensional mining to add greater power to the process (7).

There are several categories of constraints which can be used, each of which has its own characteristics and purpose. These include:

Knowledge-type constraints. This type of constraint specifies the "type of knowledge" which is to be mined, and is typically specified at the beginning of any data mining query. Some of the types of constraints which can be used include clustering, association, and classification.

Data constraints. This constraint identifies the data which is to be used in the specific data mining query. Since constraint-based mining is ideally conducted within the framework of an ad-hoc, query driven system, data constraints can be specified in a form similar to that of a SQL query.

Dimension/level constraints. Because much of the information being mined is in the form of a database or multidimensional data warehouse, it is possible to specify constraints which specify the levels or dimensions to be included in the current query.

Interestingness constraints. It would also be useful to determine what ranges of a particular variable or measure are considered to be particularly interesting and should be included in the query.

Rule constraints. It is also important to specify the specific rules which should be applied and used for a particular data mining query or application.

One application of the constraint-based approach is in the Online Analytical Mining Architecture (OLAM) developed by (7), and is designed to support the multidimensional and constraint-based mining of databases and data warehouses.

In short, constraint-based data mining is one of the developing areas which allows for the use of guiding constraints which should make for better data mining. A number of studies have been conducted in this area (4, 17, 13, 19).

PHENOMENAL DATA MINING

Phenomenal data mining is not a term for a data mining project that went extremely well. Instead, it focuses on the relationships between data and the phenomena which are inferred from the data (15). One example of this is that using receipts from cash supermarket purchases, it is possible to identify various aspects of the customers who are making these purchases. Some of these phenomena could include age, income, ethnicity, and purchasing habits. One aspect of phenomenal data mining, and in particular the goal to infer phenomena from data, is the need to have access to some facts about the relations between these data and their related phenomena. These could be included the program which examines data for phenomena, or also could be placed in a kind of knowledge base or database which can be drawn upon when doing the data mining. Part of the challenge in creating such a knowledge base involves the coding of common sense into a database, which has proved to be a difficult problem so far (14).

SUMMARY

In closing, it would not be overly optimistic to say that data mining has a bright and promising future, and that the years to come will bring many new developments, methods, and technologies. In addition, the improved integration of techniques and the application of data mining techniques can bring about the handling of new kinds of data types and applications. As the types of data and information that we have access to increases, so does the number and types of data mining which can be done. By expanding applications which can use it, integrating technologies and methods, broadening its applicability to mainstream business applications, and making programs and interfaces easier for end-users to use, it is quite possible and likely that data mining will become one of the key technology areas of the new millennium.

REFERENCES

1. Bedard, T. Merrett, and J. Han. (2001). Fundamentals of Geospatial Data Warehousing for Geographic Knowledge Discovery, H. Miller and J. Han (eds.), In *Geographic Data Mining and Knowledge Discovery*, Taylor and Francis.
2. Chakrabarti, S. (2000). Data Mining for Hypertext, *SIGKDD Explorations*, 1 (2).
3. Chakrabarti, van den Berg, and Dom. (1999) Distributed Hypertext Resource Discovery Through Examples, *Proceedings of the 25th VLDB (International Conference on Very Large Data Bases*), Edinburgh Scotland.
4. Cheung, C. Hwang, A. Fu, and J. Han. (2000). Efficient Rule-Based Attributed-Oriented Induction for Data Mining, *Journal of Intelligent Information Systems*, 15(2): 175-200.
5. Han, J. and M. Kamber. (2001). *Data Mining: Concepts and Techniques*, Morgan Kaufmann.
6. Han, J., J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, M.-C. Hsu. (2000). FreeSpan: Frequent Pattern-Projected Sequential Pattern Mining, *Proceedings KDD'00*, Boston, MA.

7. Han, J. , V. S. Lakshmanan, and R. T. Ng. (1999). Constraint-Based, Multidimensional Data Mining, *COMPUTER (special issue on Data Mining)*, 32(8): 46-50.
8. Han, J., G. Dong and Y. Yin. (1999). Efficient Mining of Partial Periodic Patterns in Time Series Database, *Proceedings International Conference on Data Engineering ICDE'99*, Sydney, Australia.
9. Kargupta, H. et al. (2000). Collective Data Mining, in *Advances in Distributed Data Mining*, Karhgupta and Chan, editors, MIT Press.
10. Kargupta, H. and A. Joshi. (2001). Data Mining To Go: Ubiquitous KDD for Mobile and Distributed Environments, Presentation, *KDD-2001*, San Francisco.
11. Kim, J. M.W. Lam, and J. Han. (2000). AIM: Approximate Intelligent Matching for Time Series Data, *Proceedings 2000 Int. Conferences on Data Warehouse and Knowledge Discovery* (DaWaK'00), Greenwich, U.K..
12. Koperski, K. and J. Han. (1995). Discovery of Spatial Association Rules in Geographic Information Databases, *Proceedings SS*D'95, Portland, Maine.
13. Lu, H., L. Feng, and J. Han. (2001). Beyond Intra-Transaction Association Analysis:Mining Multi-Dimensional Inter-Transaction Association Rules, *ACM Transactions on Information Systems*, 2001.
14. Lyons, D. And G. Tseytin, (1998). Phenomenal Data Mining and link analysis, In Jensen and Goldberg, eds., *Artificial Intelligence and Link Analysis Fall Symposium*.
15. McCarthy, J., (2000). Phenomenal Data Mining, *SIGKDD Explorations*, 1 (2), 2000. Madria et al, "Research Issues in Data Mining," In *Proceedings of Data Warehousing and Knowledge Discovery*, DaWaK '99.
16. Mizruchi, M., Mariolis, Schwartz, and Mintz. (1986). Techniques for disaggregating centrality scores in social networks, In Tuma, editor, *Sociological Methodology*, Jossey-Bass, 1986
17. Pei and J. Han. (2000). Can We Push More Constraints into Frequent Pattern Mining?, *Proceedings KDD'00*, Boston, MA.
18. Stefanovic, J. Han, and K. Koperski, (2000). Object-Based Selective Materialization for Efficient Implementation of Spatial Data Cubes, *IEEE Transactions on Knowledge and Data Engineering*, 12(6).
19. Tung, K.H., J. Han, L. V. S. Lakshmanan, and R. T. Ng. (2001). Constraint-Based Clustering in Large Databases, *Proceedings 2001 Intl. Conf. on Database Theory (ICDT'01)*, London, U.K.
20. Tung, K. H. J. Hou, and J. Han. (2001) Spatial Clustering in the Presence of Obstacles, Proc. *2001 Intl. Conf. on Data Engineering (ICDE'01*), Heidelberg, Germany.
21. Zaiane, Han, Li, and Hou. (1998). Mining Multimedia Data, *Proceedings of Meeting Of Minds CASCON '98*, Toronto Canada.
22. Zaiane, O., J. Han, and H. Zhu. (2000). Mining Recurrent Items in Multimedia with Progressive Resolution Refinement, *Proceedings International Conference on Data Engineering (ICDE 2000), San Diego CA.*
23. Zhou, D. Truffet, and J. Han. (1999). Efficient Polygon Amalgamation Methods for Spatial OLAP and Spatial Data Mining, *6th International Symposium on Spatial Databases*, SSD'99, Hong Kong.