ISSN:0975-9646

Cross Industry Survey on Data mining Applications

Ravikumar G K[1], Manjunath T. N[2], Ravindra S. Hegadi[3] ,Umesh I.M[4]

[1]Dr. MGR University, Chennai Tamilnadu, INDIA,

[2]Bharathiar University, Coimbatore Tamilnadu, INDIA,

[3]Karnatak University, Dharwad Karnataka, INDIA,

[4] R V College of Engg,Bangalore, Karnataka, INDIA,

Abstract: **Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analysis offered by data mining move beyond the analysis of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to are most likely to respond to my next promotional mailing.This paper explores on survey of the current basic technologies of data mining. Examples of profitable applications illustrate its relevance to today's business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users.**

***Keywords**: Data mining, Data warehouse, Genetic Algorithms, Association, Clustering, Classification*

1 INTRODUCTION

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond exposition data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

(1)Massive data collection (2) Powerful multiprocessor computers and (3) Data mining algorithms.

Commercial databases are growing at unprecedented rates. A recent META Group survey of data warehouse projects found that 19 percent of respondents are beyond the 50 gigabyte level, while 59 percent expect to be there by second quarter of 1996. In some industries, such as retail, these numbers can be much larger. The accompanying need for improved computational engines can now be met in a cost-effective manner with parallel multiprocessor computer technology. Data mining algorithms embody techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods. In the evolution from business data to business information, each new step has built upon the previous one. From the user's point of view, the four steps listed in Table-1 were revolutionary because they allowed new business questions to be answered accurately and quickly.

Evolutionary Step	**Business Question**	**Enabling Technologies**
Data Collection (1960s)	"What was my total revenue in the last five years"	Computers, tapes, disks
Data Access (1980s)	"What were unit sales in England last march?"	Relational databases, Structured Query Language, ODBC
Data warehousing& Decision Support (1990s)	"What were unit sales in England last March? Drill down to Boston."	On-line analytic processing, multidimensional databases, data warehouses
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases

Table-1 Steps in Evolution of Data Mining

The core components of data mining technology have been under development for decades, in research areas such as statistics, artificial intelligence, and machine learning. Today, the maturity of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments.META Group Application Development Strategies: "Data Mining for Data Warehouses: Uncovering Hidden Patterns".

1.1 Data Mining in Diversified fields

Data mining derives its name from the similarities between searching for valuable business information in a large database for example, finding linked products in gigabytes of store scanner data and mining a mountain for a vein of valuable ore. Both processes require either shifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities.

1.1.1 Automated prediction of trends and behaviors

Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

1.1.2 Automated discovery of previously unknown patterns

Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions.

1.1.3 Databases depth and breadth

More Columns. Analysts must often limit the number of variables they examine when doing hands-on analysis due to time constraints. Yet variables that are discarded because they seem unimportant may carry information about unknown patterns. High performance data mining allows users to explore the full depth of a database, without preselecting a subset of variables.

More Rows. Larger samples yield lower estimation errors and variance, and allow users to make inferences about small but important segments of a population.

A recent Gartner Group Advanced Technology Research Note listed data mining and artificial intelligence at the top of the five key technology areas that "will clearly have a major impact across a wide range of industries within the next 3 to 5 years."* Gartner also listed parallel architectures and data mining as two of the top 10 new technologies in which companies will invest during the next 5 years. According to a recent Gartner HPC Research Note, "With the rapid advance in data capture, transmission and storage, large-systems users will increasingly need to implement new and innovative ways to mine the after-market value of their vast stores of detail data, employing MPP [massively parallel processing] systems to create new sources of business advantage (0.9 probability)."**

2 TECHNIQUES USED IN DATA MINING

2.1 Artificial Neural Networks

Non-linear predictive models that learn through training and resemble biological neural networks in structure.

2.2 Decision Trees

Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

2.3 Genetic Algorithms

Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

2.4 Nearest Neighbor Method

A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k is greater than or equal to 1). Sometimes called the k-nearest neighbor technique.

2.5 Rule Induction

The extraction of useful if-then rules from data based on statistical significance.

Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms.

3 HOW DATA MINING WORKS

How exactly is data mining able to tell you important things that you didn't know or what is going to happen next? The technique that is used to perform these feats in data mining is called modeling. Modeling is simply the act of building a model in one situation where you know the answer and then applying it to another situation that you don't. For instance, if you were looking for a sunken Spanish galleon on the high seas the first thing you might do is to research the times when Spanish treasure had been found by others in the past. You might note that these ships often tend to be found off the coast of Bermuda and that there are certain characteristics to the ocean currents, and certain routes that have likely been taken by the ship's captains in that era. You note these similarities and build a model that includes the characteristics that are common to the locations of these sunken treasures. With these models in hand you sail off looking for treasure where your model indicates it most likely

might be given a similar situation in the past. Hopefully, if you've got a good model, you find your treasure. This act of model building is thus something that people have been doing for a long time, certainly before the advent of computers or data mining technology. What happens on computers, however, is not much different than the way people build models.

Computers are loaded up with lots of information about a variety of situations where an answer is known and then the data mining software on the computer must run through that data and distill the characteristics of the data that should go into the model. Once the model is built it can then be used in similar situations where you don't know the answer. For example, say that you are the director of marketing for a telecommunications company and you'd like to acquire some new long distance phone customers. You could just randomly go out and mail coupons to the general population - just as you could randomly sail the seas looking for sunken treasure. In neither case would you achieve the results you desired and of course you have the opportunity to do much better than random you could use your business experience stored in your database to build a model.As the marketing director you have access to a lot of information about all of your customers: their age, sex, credit history and long distance calling usage. The good news is that you also have a lot of information about your prospective customers: their age, sex, credit history etc. Your problem is that you don't know the long distance calling usage of these prospects (since they are most likely now customers of your competition). You'd like to concentrate on those prospects that have large amounts of long distance usage. You can step in Evolution of Data Mining accomplish this by building a model.

Table-2 illustrates the data used for building a model for new customer prospecting in a data warehouse.

Type of Info	Customer	Prospects
General Information (eg:Demographic data)	Known	Known
Proprietary information	Known	Target

Table-2 Data mining for prospecting

The goal in prospecting is to make some calculated guesses about the information in the lower right hand quadrant based on the model that we build going from Customer General Information to Customer proprietary Information. For instance, a simple model for a telecommunications company might be: 98 percent of my customers who make more than $60,000 per year spend more than $80 per month on long distance. This model could then be applied to the prospect data to try to tell something about the proprietary information that this telecommunications company does not currently have access to. With this model in hand, new customers can be selectively targeted. Test marketing is an excellent source of data for this kind of modeling. Mining the results of a test market representing a broad but relatively small sample of prospects can provide a foundation for identifying good prospects in the overall market. Table-3 shows another common scenario for building models: predict what is going to happen in the future.

Type of info	Yesterday	Today	Tomorrow
Static Information and current plans (eg:Demographic data, marketing plans)	Known	Known	Known
Dynamic information(eg:Customer Transaction)	Known	Known	Target

Table-3 Data Mining for Prediction

If someone told you that he had a model that could predict customer usage how would you know if he really had a good model? The first thing you might try would be to ask him to apply his model to your customer base - where you already knew the answer. With data mining, the best way to accomplish this is by setting aside some of your data in a vault to isolate it from the mining process. Once the mining is complete, the results can be tested against the data held in the vault to confirm the model's validity. If the model works, its observations should hold for the vaulted data.

4 AN ARCHITECTURE FOR DATA MINING

To best apply these advanced techniques, they must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data warehouse can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on. Figure-1 illustrates architecture for advanced analysis in a large data warehouse.

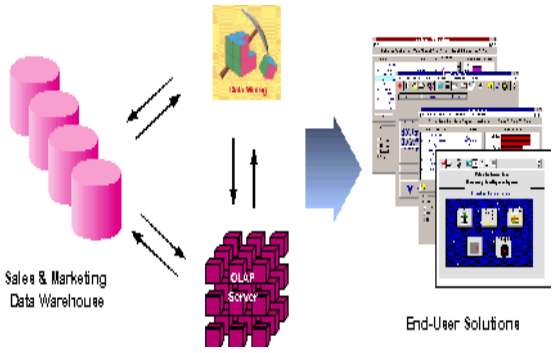

Figure-1 Integrated Data Mining Architecture

The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting. This warehouse can be implemented in a variety of relational database systems: Sybase, Oracle, Redbrick, and so on, and should be optimized

for flexible and fast data access. An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model to be applied when navigating the data warehouse. The multidimensional structures allow the user to analyze the data as they want to view their business summarizing by product line, region, and other key perspectives of their business. The Data Mining Server must be integrated with the data warehouse and the OLAP server to embed ROI-focused business analysis directly into this infrastructure. An advanced, process-centric metadata template defines the data mining objectives for specific business issues like campaign management, prospecting, and promotion optimization. Integration with the data warehouse enables operational decisions to be directly implemented and tracked. As the warehouse grows with new decisions and results, an organization can continually mine the best practices and apply them to future decisions. This design represents a fundamental shift from conventional decision support systems. Rather than simply delivering data to the end user through query and reporting software, the Advanced Analysis Server applies users' business models directly to the warehouse and returns a proactive analysis of the most relevant information. These results enhance the metadata in the OLAP Server by providing a dynamic metadata layer that represents a distilled view of the data. Reporting, visualization, and other analysis tools can then be applied to plan future actions and confirm the impact of those plans.

4.1 Profitable Applications

A wide range of companies have deployed successful applications of data mining. While early adopters of this technology have tended to be in information-intensive industries such as financial services and direct mail marketing, the technology is applicable to any company looking to leverage a large data warehouse to better manage their customer relationships. Two critical factors for success with data mining are: a large, well-integrated data warehouse and a well-defined understanding of the business process within which data mining is to be applied (such as customer prospecting, retention, campaign management, and so on).

Some successful application areas include:

A pharmaceutical company can analyze its recent sales force activity and their results to improve targeting of high-value physicians and determine which marketing activities will have the greatest impact in the next few months. The data needs to include competitor market activity as well as information about the local health care systems. The results can be distributed to the sales force via a wide-area network that enables the representatives to review the recommendations from the perspective of the key attributes in the decision process. The ongoing, dynamic analysis of the data warehouse allows best practices from throughout the organization to be applied in specific sales situations.

A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product. Using a small test mailing, the attributes of customers with an affinity for the product can be identified. Recent projects have indicated more than a 20-fold decrease in costs for targeted mailing campaigns over conventional approaches. A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. Using data mining to analyze its own customer experience, this company can build a unique segmentation identifying the attributes of high-value prospects. Applying this segmentation to a general business database such as those provided by Dun & Bradstreet can yield a prioritized list of prospects by region.

A large consumer package goods company can apply data mining to improve its sales process to retailers. Data from consumer panels, shipments, and competitor activity can be applied to understand the reasons for brand and store switching. Through this analysis, the manufacturer can select promotional strategies that best reach their target customer segments. Each of these examples has a clear common ground. They leverage the knowledge about customers implicit in a data warehouse to reduce costs and improve the value of customer relationships. These organizations can now focus their efforts on the most important (profitable) customers and prospects, and design targeted marketing strategies to best reach them.

5 CONCLUSION

Comprehensive data warehouses that integrate operational data with customer, supplier, and market information have resulted in an explosion of information. Competition requires timely and sophisticated analysis on an integrated view of the data. However, there is a growing gap between more powerful storage and retrieval systems and the users' ability to effectively analyze and act on the information they contain. Both relational and OLAP technologies have tremendous capabilities for navigating massive data warehouses, but brute force navigation of data is not enough. A new technological leap is needed to structure and prioritize information for specific end-user problems. The data mining tools can make this leap. Quantifiable business benefits have been proven through the integration of data mining with current information systems, and new products are on the horizon that will bring this integration to an even wider audience of users.

ACKNOWLEDGEMENTS

This paper is prepared through exhaustive discussions and T-cons with Subject Matter Experts (SME), data warehouse groups, Data Quality Experts of various organizations in India and abroad. The authors gratefully acknowledge the time spend in this discussions provided by Mr.Shahzad, SME, CSC USA, Mr. Parswanath Project Manager (Data Warehouse Wing). Wipro Technologies, India. Mr. Govardhan (Architect) IBM India Pvt Ltd, Mr. Arun Kumar Data *Architect* KPIT Cummins India.

REFERENCES

[1] Dianhui Wang, Yong-Soo Kim, Seok Cheon Park, Chul Soo Lee and Yoon Kyung Han "Learning Based Neural Similarity Metrics for Multimedia Data Mining ", Soft Computing, Volume 11, Number 4, February 2007, pp. 335-340.

[2] Bhavani Thuraisingham, Managing and Mining Multimedia, Databases, Published by CRC Press, 2001

[3] Sanjeevkumar R. Jadhav, and Praveen Kumar Kumbargoudar, "Multimedia Data Mining in Digital Libraries: Standards and Features" in Proc. READIT-2007, p. 54.

[4] Shu-Ching Chen, Mei-Ling Shyu, Chengcui Zhang, and Jeff Strickrott, "Multimedia Data Mining for Traffic Video Sequences," Proceedings of the Second International Workshop on Multimedia data Mining MDM/KDD'2001), in conjunction with the Seventh ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 78-85, August 26, 2001, San Francisco, CA, USA.

[5] Valery A. Petrushin and Latifur Khan, "Multimedia Data Mining and Knowledge Discovery", 2007 - London: Springer-Verlag, pp. 3-17

[6] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, Multimedia Mining, WSEAS Transactions on Systems, Issue 10, Volume 3, December 2004, pp. 3263-3268

[7] Sanjiv Purba "Data Management Handbook" Published by CRC Press, 1999

[8] Bhavani M. Thuraisingham, "Data Management Systems: Evolution and Interoperation", Published by CRC Press, 1997

[9] Jiawei Han, Micheline Kamber "Data Mining: Concepts and Techniques" Published by Morgan Kaufmann, 2001

[10] Sanjeevkumar R. Jadhav, and Praveen Kumar Kumbargoudar, Multimedia Data Mining in Digital Libraries: Standards and Features, ACVIT- 07, Dr. Babasaheb Ambedkar MarathWada University, Aurangabad,MS-India

[11] Mori Y, Takahashi H, Oka R. Image-to-word transformation based on dividing and vector quantizing images with words. In: MISRM'99 First International Workshop on Multimedia Intelligent Stotage and Retrievel mnagement, 1999

[12] Ordenoz C, Omiecinski E. Discovering association rules based on image content. In: ADL '99: Proceedings of the IEEE Forum on Research and Technology Advances in Digital libraries. Washington, DC: IEEE Computer Society; 1999, p.38.

[13] Tseng, V.S.; Ming-Hsiang Wang; Ja-Hwung Su, A New Method for Image Classification by Using Multilevel Association Rules, Data Engineering Workshops, 05-08 April 2005 Page(s): 1180 – 1180 .

[14] Ankur M. Teredesai, Muhammad A. Ahmad, Juveria Kanodia and Roger S. Gaborski, "Knowledge and Information Systems", Volume 10, August, 2006, Springer London

[15] Mittal, Ankush An overview of multimedia content-based retrieval strategies, Publication: Informatica, October 1 2006

[16] J. You, J. Liu, (PRC), L. Li (Australia), and K.H. Cheung (PRC), on data mining and warehousing for multimedia information retrieval,From Proceeding (365) Artificial and Computational Intelligence – 2002

[17] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, Multimedia Mining-2004. 'Multimedia mining', WSEAS Transactions on Systems,. Vol. 3, No. 10, pp.3263–3268.

[18] Manjunath T.N, Ravindra S Hegadi, Ravikumar G K."Analysis of Data Quality Aspects in DataWarehouse Systems", International Journal of Computer Science and Information Technologies, Vol. 2 (1), 2010, 477-485

[19] Yu H, Wolf, Scenic classification methods for image and video databases. In In SPIE International Conference on Digital Image Storage and Archiving Systems, Vol. 2606, 1995, pp. 363-371.

[21] Manjunath T.N, Ravindra S Hegadi, Ravikumar G K." A Survey on Multimedia Data Mining and Its Relevance Today", International journal of Computer Science and Network Security. Vol. 10 No. 11 pp. 165-170.

Ravikumar GK. received his Bachelor's degree from Siddaganga Institute of Technology, Tumkur (Bangalore University) during the year 1996 and M. Tech in Systems Analysis and Computer Application from Karnataka Regional Engineering College Surthakal (NITK) during the year 2000. He is currently working towards his PhD degree in the Area of Data mining . He has published several papers in International and national level conferences. He is having around 14 years of Professional experienced which includes Software Industry and teaching experience. His area of interests are Data Warehouse & Business Intelligence, multimedia and Databases.

Manjunath T N. received his Bachelor's Degree in computer Science and Engineering from SJC Institute of Technology, Chickballapur, Karnataka, India during the year 2001 and M. Tech in computer Science and Engineering from Jawaharlal Nehru National College of Engineering, Shimoga, Karnataka, India during the year 2004. Currently pursing Ph.D degree in Bharathiar University, Coimbatore. He is having total 10 years of Industry and teaching experience. His areas of interests are Data Warehouse & Business Intelligence, multimedia and Databases. He has published and presented papers in journals, international and national level conferences.

Dr.Ravindra S Hegadi received his Master of Computer Applications (MCA) & M.Phil and Doctorate of Philosophy Ph.D. in year 2007 in computer science from Gurbarga University, Karnataka; He is having 15 years of Experience. He has visited overseas to various universities as SME.His area of interests are Image Mining, Image Processing and Databases and business intelligence. He has published and presented papers in journals, international and national level conferences.

Umesh.I.M. received his Master of Science (MSc) & M.Phil in year 2007 in computer science from Bharathidasan University, Tamilnadu,He is working in R V College of Engineering,Bangalore,Karnataka,India.He is having 10 years of Experience.His area of interests are Image Mining, Image Processing and Databases and business intelligence. He has published and presented papers in journals, international and national level conferences