Data Mining and Knowledge Management

Vinay Singh[1], Kaushal Kumar[2]

[1,2]Research Scholar Department of Mechanical Engineering, GJUS&T, Hisar, Haryana, India

---

Abstract - *Rapid increases in technological and informational systems have led businesses to collect customer data in huge databases. Data mining is the process involving analyzing, searching data to make it useful for human use. Large amount of data is modeled, selected and explored in order to determine comprehensible information. This article represents data mining tools used to understand the data mining process. Also, data mining enablers as well as barriers are also described to make the subject more understandable.*

Key Words*: *Data mining, Data mining models, Knowledge Management, Enablers, Barriers

1. INTRODUCTION

Data mining (DM) is the process of trawling through data to find previously unknown relationships among the data that are interesting to the user of the data (Hand, 1998). Data Mining has been an established field (Fayyad et al., 1996; Chen and Liu, 2005; Wang, 2005). Data mining is the process of searching and analyzing data in order to find implicit, but potentially useful, information (M.J.A. Berry et al, 1997). It involves selecting, exploring and modeling large amounts of data to uncover previously unknown patterns, and ultimately comprehensible information, from large databases (Shaw et al, 2001). Data mining uses a broad family of computational methods that include statistical analysis, decision trees, neural networks, rule induction and refinement, and graphic visualization (Brachman, 1996). Also, Data Mining techniques should be carefully understood and applied by the frontline users (Hall, 2004; Violino, 2004; King, 2005). Data mining allows a search, for valuable information, in large volumes of data. The explosive growth in databases has created a need to develop technologies that use information and knowledge intelligently (Weiss & Indurkhya, 1998). According to Rubenking (2001), "data mining is the process of automatically extracting useful information and relationships from immense quantities of data. In its purest form, data mining doesn't involve looking for specific information. Data mining is an interdisciplinary field that combines artificial intelligence, database management, data visualization, machine learning, mathematic algorithms, and statistics. Data mining, also known as knowledge discovery in databases (KDD) (Chen, Han, & Yu, 1996; Fayyad, Piatetsky-Shapiro, & Smyth, 1996a), is a rapidly emerging field. This technology provides different methodologies for decision-making, problem solving, analysis, planning, diagnosis, detection, integration, prevention, learning, and innovation. Data Mining was defined by Turban, Aronson, Liang, and Sharda (2007) as a process that uses statistical, mathematical, artificial intelligence and machine-learning techniques to extract and identify useful information and subsequently gain knowledge from large databases.

2. Data Mining and Knowledge Management

Knowledge discovery and learning is an iterative process that extends the collection of data mining techniques into a knowledge management framework (Michael J. Shaw, 2001). Higher education will find larger and wider applications for data mining than its counterpart in the business sector, because higher education institutions carry three that data mining intensive duties: scientific research that relates to the creation of knowledge, teaching that concerns with the transmission of knowledge, and institutional research that pertains to the use of knowledge for decision making. All the above tasks are well within the boundaries of Knowledge Management, which drives the need for better and faster decision making tools and methods (Luan Jing, 2005). Owing to its strength, Data Mining is known as a powerful Business Intelligence tool for knowledge discovery (Chen and Liu, 2005). The process of Data Mining is a Knowledge Management process because it involves human knowledge (Brachman et al., 1996).

Several authors have also written about the factors behind the dawn of data mining. For instance, Therling (1995) identified three reasons: The ease of data collection and storage, the computing power of modern processors, and the need for fast and real time data mining. Yet, one important reason absent from these is the growing interest in Knowledge Management.

3. Data Mining Tools

a) Web-based software tools: To meet the competitive global challenges, the firm's knowledge workers require improved tools for understanding the changing markets and customer requirements. Historically, forecasting tools were the primary business insight generation tools used to analyze the competitive landscape (D.N. Clark, 1992). The business objective for using these insight-generation tools was to help knowledge workers predict the future of a given market segment or the success of a particular product line. These forecasting tools aided in reducing decision uncertainty by providing a degree of confidence to those decisions related to the success of market segments or product lines (G.J. Browne et al, 1997).

b) Business model: The business models utilized by the knowledge workers can be categorized into assessment models, tactical models, and strategic models. These business models assist the knowledge worker in making sense of the competitive landscape and in providing the knowledge workers the needed focus (J.F. Courtney, 2001),(B.T. Gale,1994).

c) The WEKA Data Mining Software: The Waikato Environment for Knowledge Analysis (WEKA) came about through the perceived need for a unified workbench that would allow researchers easy access to state-of the- art techniques in machine learning. It was envisioned that WEKA would not only provide a toolbox of learning algorithms, but also a framework inside which researchers could implement new algorithms without having to be concerned with supporting infrastructure for data manipulation and scheme evaluation. (Mark Hall)Nowadays, WEKA is recognized as a landmark system in data mining and machine learning (G. Piatetsky, 2005).

d) Artificial neural network (ANN): Artificial neural network (ANN) has popularity in solving several problems and technical problems that involve prediction, and have a wide ranging usage area is one of the most important data mining techniques (K. Usha Rani, 2011). An artificial neural network (ANN) is a computational model based on biological neural networks and consists of an interconnected group of artificial neurons. It can be treated as non-linear statistical data modelling tools that can be used to model complex relationships between inputs and outputs or to find patterns in data (M. Kamrunnahar, 2010).

4. Data Mining Models

a) Association: Association aims to establishing relationships between items which exist together in a given record (Ahmed, 2004; Jiao, Zhang, & Helander, 2006; Mitra et al., 2002). Market basket analysis and cross selling programs are typical examples for which association modelling is usually adopted. Common tools for association modelling are statistics and algorithms.

b) Classification: Classification is one of the most common learning models in data mining (Ahmed, 2004; Berry & Linoff, 2004; Carrier & Povel, 2003). It aims at building a model to predict future customer behaviours through classifying database records into a number of predefined classes based on certain criteria (Ahmed, 2004; Berson et al., 2000; Chen, Hsu, & Chou, 2003; Mitra et al., 2002). Common tools used for classification are neural networks, decision trees and if then-else rules.

c) Clustering: Clustering is the task of segmenting a heterogeneous population into a number of more homogenous clusters (Ahmed, 2004; Berry & Linoff, 2004; Carrier & Povel, 2003; Mitra et al., 2002). It is different to classification in that clusters are unknown at the time the algorithm starts. Common tools for clustering include neural networks and discrimination analysis.

d) Forecasting: Forecasting estimates the future value based on a record's patterns. It deals with continuously valued outcomes (Ahmed, 2004; Berry & Linoff, 2004). It relates to modelling and the logical relationships of the model at some time in the future. Demand forecast is a typical example of a forecasting model. Common tools for forecasting include neural networks and survival analysis.

e) Regression: Regression is a kind of statistical estimation technique used to map each data object to a real value provide prediction value (Carrier & Povel, 2003; Mitra et al., 2002). Uses of regression include curve fitting, prediction (including forecasting), modeling of causal relationships, and testing scientific hypotheses about relationships between variables. Common tools for regression include linear regression and logistic regression

f) Sequence discovery: Sequence discovery is the identification of associations or patterns over time (Berson et al., 2000; Carrier & Povel, 2003; Mitra et al., 2002). Its goal is to model the states of the process generating the sequence or to extract and report deviation and trends over time (Mitra et al., 2002). Common tools for sequence discovery are statistics and set theory.

g) Visualization: Visualization refers to the presentation of data so that users can view complex patterns (Shaw et al., 2001). It is used in conjunction with other data mining models to provide a clearer understanding of the discovered patterns or relationships (Turban et al., 2007). Examples of visualization model are 3D graphs, ''Hygraphs'' and ''SeeNet'' (Shaw et al., 2001).

h) Summarization: It involves the finding a compact description for a subset of data, e.g., the derivation of summary or association rules and the use of multivariate visualization techniques (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

i)Dependency Modeling: This modeling includes finding a model which describes significant dependencies between variables (e.g., learning of belief networks)(Fayyad, Piatetsky-Shapiro, & Smyth,1996).

j) Change and Deviation Detection: This process contributes discovering the most significant changes in the data from previously measured or normative values). (Fayyad, Piatetsky- Shapiro, & Smyth, 1996).

Table -1: Enablers of Data Mining

Enablers of Data Mining			
S. No	Barriers	Description	References
1	Information Technology	The advent of information technology has transformed the way marketing is done and how companies manage information	Kumar et al. 2014, Kumar et al. 2014

		about their customers.	
2	Internet and the World Wide Web	The Internet and the World Wide Web have made the process of collecting data easier, adding to the volume of data available to businesses	Michael J. Shaw, C Subramaniam Gek Woo Tan Michael E. Welge , 2001
3	Effective customer relationship management	Data mining tools can help uncover the hidden knowledge and understand customer better, while a systematic knowledge management effort can channel the knowledge into effective marketing strategies.	D. Peppers, M.Rogers,1997
4	Developments in database processing	Developments in database processing, data warehousing, machine learning and knowledge management have contributed greatly to our understanding of the data mining process.	K.C.C.C. Chan, A.K.C. Wong, 1991, M. Holsheimer, M.L. Kersten, A.P.J. M. Siebes, 1996,
5	Data warehousing	Developments in database processing, data warehousing, machine learning and knowledge management have contributed greatly to our understanding of the data mining process.	W. Inmon, 1996,
6	Machine learning	Developments in database processing, data warehousing, machine learning and knowledge management have contributed greatly to our understanding of the data mining process.	C.W. Holsapple, R. Pakath, V.S. Jacob, J.S. Zaveri, 1993,
7	Knowledge management	Developments in database processing, data warehousing, machine learning and knowledge management have contributed greatly to our understanding of the data mining process.	D.M. Amidon, 1998,H. Holtz, 1992, M.C. Rumizen,1998
8	Advances in computer hardware and software	Although, data mining tools have been available for a long time, the advances in computer hardware and software, particularly exploratory tools like data visualization and neural networks, have made data mining more attractive and practical.	U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, 1996
9	Strategic business initiatives	These initiatives support applications that help increase efficiency and improve effectiveness of the firm to moving massive paper-based information sources into electronic form, to facilitating data mining and insight generation.	L. Bransten , 1999
10	Knowledge Discovery in Databases	It is the process of using the database along with any required selection, preprocessing, sub sampling, and transformations of it; to apply data mining methods (algorithms)	Fayyad, Piatetsky-Shapiro, & Smyth (1996).

		to enumerate patterns from it.	

Table -2: Barriers of Data Mining

Barriers of Data Mining			
S. No	Barriers	Description	References
1	Intense competition and increased choices	The intense competition and increased choices available for customers in market have created new pressures on marketing decision-makers and there has emerged a need to manage customers in a long-term relationship	D.Peppers, M.Rogers, 1999
2	Technical complexity issues	Technical complexity issues, lack of senior management focus, inflexibility of the software tools, and difficulty in assessing benefits provided to the firm are the main reasons to explain the relatively low implementation success rate.	John H. Heinrichs , Jeen-Su Lim, 2003
3	Lack of senior management focus	Lack of senior management focus produces relatively low implementation success rate and the relatively low satisfaction ratings from these projects	G. Bassellier, B.H. Reich, I. Benbasat, 2001
4	Inflexibility of the software tools	Inflexibility of the software tools hinders in proper implementation success rate.	L. Bransten,1999
5	Difficulty in assessing benefits provided to the firm.	Tool industry segment continues to experience a dramatic 40% compounded annual sales growth rate because of the difficulty in assessing the benefits provided to the firm	John H. Heinrichs , Jeen-Su Lim, 2003
6	Larger databases	Databases with hundreds of fields and tables, millions of records are quite common. Methods for dealing with large data volumes include more efficient algorithms, sampling, approximation methods, and massively parallel processing	Agrawal et al. 1996, Holsheimer et al. 1996
7	High dimensionality	There is often a very large number of records in the database, and can also be a very large number of fields (attributes, variables) so that the dimensionality of the problem is high.	Fayyad, Piatetsky-Shapiro, & Smyth 1996
8	Overfitting	When the algorithm searches for the best parameters for one particular model using a limited set of data, it may resulting in poor performance of the model on test data. Possible solutions include cross-validation, regularization, and other sophisticated statistical strategies.	Fayyad, Piatetsky-Shapiro, & Smyth, 1996

9	Changing data and knowledge	Rapidly changing (non-stationary) data may make previously discovered patterns invalid. In addition, the variables measured in a given application database may be modified, deleted, or augmented with new measurements over time.	Mannila, Toivonen, & Verkamo 1995; Agrawal & Psaila 1995, Matheus, Piatetsky-Shapiro, and Mc-Neill, 1996
10	Missing and noisy data	Important attributes may be missing if the database was not designed with discovery in mind. Possible solutions include more sophisticated statistical strategies to identify hidden variables and dependencies	Heckerman 1996; Smyth et al. 1996
11	Complex relationships between fields:	Hierarchically structured attributes or values, relations between attributes, and more sophisticated means for representing knowledge about the contents of a database will require algorithms that can effectively utilize such information.	Djoko, Cook, & Holder 1995; Dzeroski 1996
12	User interaction and prior knowledge	Many current Knowledge Discovery in Database methods and tools are not truly interactive and cannot easily incorporate prior knowledge about a problem except in simple ways	Cheeseman, P. 1990
13	Integration with other systems	A stand-alone discovery system may not be very useful. Typical integration issues include integration with a DBMS, integration with spreadsheets and visualization tools, and accommodating real-time sensor readings.	Simoudis, LivezeyKerber 1995 and Stolorz et al. 1995

7. CONCLUSIONS

The concept of data mining has been explained in the paper. Data mining has been proved as a better tool for knowledge discovery. Several data mining tools have been discussed in the literature primarily includes: Web based software tool, WEKA software tool and artificial neural network. Several data mining models have also been discussed. Also, the enablers and barriers of data mining process along with their descriptions have also been discussed. Knowledge management and data mining is also correlated in the article. This article will help the academicians to get in depth knowledge of the concept of data mining.

REFERENCES

Agrawal, Ft. and Psaila, G. 1995. Active Data Mining, In Proceedings of KDD-95: First International Conference on Knowledge Discovery and Data Mining, pp. 3-8, Menlo Park, CA: AAAI Press.

Agrawal R. Mannila, H., Srikant, Ft., Toivonen, H., and Verkamo, I. 1996. Fast Discovery of Association Rules in AKDDM, AAAI/MIT Press, 307 328.

Ahmed, S. R. (2004). Applications of data mining in retail business. Information Technology: Coding and Computing, 2, 455–459.

Amidon, D. M. (1998). Blueprint for 21st century innovation management.Journal of Knowledge Management, 2(1), 23-31.

B.T. Gale, Managing Customer Value: Creating Quality and Service That Customers Can See, The Free Press, New York, NY, 1994.

Bassellier, G., Benbasat, I., & Reich, B. H. (2003). The influence of business managers' IT competence on championing IT. Information Systems Research, 14(4), 317-336.

Berry, M. J. A., & Linoff, G. S. (2004). Data mining techniques second edition – for marketing, sales, and customer relationship management, Wiley.

Berson, A., Smith, S., & Thearling, K. (2000).Building data mining applications for CRM. McGraw-Hill.

Brachman, R.J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G. and Simoudis, E. (1996), "Mining business databases", Communications of the ACM, Vol. 39 No. 11, pp. 42-8.

Bassellier, G., Benbasat, I., & Reich, B. H. (2003). The influence of business managers' IT competence on championing IT. Information Systems Research, 14(4), 317-336.

Chan, K. C. C., & Wong, A. K. C. (1991). Knowledge Discovery in Databases

Carrier, C. G., & Povel, O.(2003). Characterising data mining software.Intelligent Data Analysis, 7, 181–192.

Cheeseman, P.1990.On Finding the Most Probable Model. In Computational Models o] Scientific Discovery and Theory Formation, Shrager, J. and Langley P. (eds). Los Gatos, CA: Morgan Kaufinann, 73 95

Chen, M. S., Han, J., &Yu, P. S. (1996). Data mining: an overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering, 8(6), 866–883

Chen, Y. L., Hsu, C. L., & Chou, S. C. (2003) Constructing a multi-valued and multilabeled decision tree. Expert Systems with Applications, 25, 199–209.

Chen, S.Y. and Liu, X. (2005), "Data mining from 1994 to 2004: an application-oriented review", International Journal of Business Intelligence and Data Mining, Vol. 1 No. 1, pp. 4-11.

Djoko, S., Cook, D., and Holder, I, 1995. Analyzing the Benefits of Domain Knowledge in Substructure Discovery, in Proceedings o] KDD-95: First International Conference on Knowledge Discovery and Data Mining, Menlo Park, CA: The AAAI Press.

D.N. Clark, (1992), A literature analysis of the use of management science tools in strategic planning, Journal of Operations Research Society 43 (9) 859–870.

D. Peppers, M. Rogers, Enterprise One to One: Tools for Competing in the Interactive Age, Doubleday, New York, 1997.

Dzeroski, S. 1996.Inductive Logic Programming for Knowledge Discovery in Databases, in A KDDM, AAAI/MIT Press.

F'ayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. 1996. From Data Mining to Knowledge Discovery: An Overview, in AI (DDM, AAAI/MIT Press, pp. 1-30.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From data mining to knowledge discovery: an overview.In Advances in knowledge discovery and data mining (pp.1-34). California: American Association for Artificial Intelligence

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996), "The KDD process for extracting useful knowledge from volumes of data", Communications of the ACM, Vol.39 No. 11, pp. 7-34.

G.J. Browne, S.P. Curley, G.P. Benson, Evoking information in probability assessment: knowledge maps and reasoning-based directed questions, Management Science 43 (1) (1997) 1 –14.

G. Piatetsky-Shapiro. KD nuggets news on SIGKDD service award. http://www.kdnuggets.com/news/2005/n13/2i.html, 2005.

Hall, M. (2004), "Doubtful BI", Computerworld, Vol. 38 No. 25, p. 45.

Hand, D.J. (1998), "Data mining: statistics and more?" The American Statistician, Vol. 52 No. 2, pp. 112-8.

Holsheimer, M., Kersten, M.L, Mannila, H., and Toivonan,H. 1996. Data Surveyor: Searching the Nuggets in Parallel, in AKDDM, AAAI/MIT Press.

Heckerman, D. 1996. Bayesian Networks for Knowledge Discovery, in ilKDDM, AAAI/MIT Press, 273-306.

Heinrichs, J. H., & Lim, J. S. (2003). Integrating web-based data mining tools with business models for knowledge management. Decision Support Systems, 35(1), 103-112.

Holsapple, C. W., Pakath, R., Jacob, V. S., & Zaveri, J. S. (1993). Learning by problem processors: adaptive decision support systems. Decision Support Systems, 10(2), 85-108.

Inmon, W.H. (1996), Building the Data Warehouse, John Wiley & Sons Inc., New York, NY.

J.F. Courtney, Decision making and knowledge management in inquiring organizations: toward a new decision-making paradigm for DSS, Decision Support Systems 31 (2001) 17– 38.

Jiao, J. R., Zhang, Y., & Helander, M. (2006).A Kansei mining system for affective design. Expert Systems with Applications, 30, 658–673.

K. Usha Rani, International Journal of Data Mining and Knowledge Management Process (IJDKP) 1 (5) (2011).

King, J. (2005), "Better decisions", Computerworld, Vol. 39 No. 38, pp. 48-9.

Kersten, M. L., Siebes, A. P. J. M., Holsheimer, M., & Kwakkel, F. (1997). Research and business challenges in data mining technology. InDatenbanksysteme in Büro, Technik und Wissenschaft (pp. 1-16). Springer Berlin Heidelberg.

Kumar, S., Singh, V., & Haleem, A. (2015). Critical success factors of knowledge management: modelling and comparison using various techniques. International Journal of Industrial and Systems Engineering, 21(2), 180-206.

Kumar, S., Singh, V., & Haleem, A. (2014). Knowledge management–enablers and barriers: a questionnaire–based study. International Journal of Knowledge Engineering and Data Mining, 3(1), 31-57.

Luan Jing, 2005, "Data Mining and Knowledge Management in Higher Education Potential Applications", Annual Forum for the Association, for Institutional Research (42nd, Toronto, Ontario, Canada,June 2-5, 2002).

L. Bransten, Looking for patterns, The Wall Street Journal (June 21, 1999) R16

Mannila, tt., Toivonen, H. and Verkamo, A.I. 1995. Discovering Frequent Episodes in Sequences, In Proceedings of KDD-95: First International Conference on Knowledge Discovery and Data Mining, pp. 210-215, Menlo Park, CA: AAAI Press.

Matheus, C. J., Piatetsky-Shapiro, G., & McNeill, D. (1996). 20 Selecting and Reporting What is Interesting: The KEFIR Application to Healthcare Data.

M. Kamrunnahar, M. Urquidi-Macdonald, Corros. Sci. 52 (2010) 669–677

Michael J. Shaw, Chandra Sekar Subramaniam, Gek Woo Tan and Michael E. Welge, 2001. Knowledge Management and Data Mining for Marketing, Decision Support Systems, 31: 127-137

Mitra, S., Pal, S. K., & Mitra, P. (2002). Data mining in soft computing framework: A survey. IEEE Transactions on Neural Networks, 13, 3–14.

M.J.A. Berry, G. Linoff, Data Mining Techniques for Marketing, Sales, and Customer Support, Wiley, New York, 1997.

Maurer, F., & Holz, H. (1999, October). Process-oriented knowledge management for learning software organizations. In Proceedings of 12th Knowledge Acquisition For Knowledge-Based Systems Workshop.

Peppers, D., Rogers, M., & Dorf, B. (1999). Is your company ready for oneto-one marketing? Harvard Business Review, 77(1), 101 – 119.

Rubenking, N. (2001) Hidden Messages. PC Magazine. May 22, 2001.

Rumizen, M. C. (1998). Report on the second comparative study of knowledge creation conference. Journal of Knowledge Management, 2(1), 77-82.

Shaw, M. J., Subramaniam, C., Tan, G. W., & Welge, M. E. (2001).Knowledge management and data mining for marketing. Decision Support Systems, 31, 127–137.

Simoudis, E., Livezey, B., and Kerber, R. 1995. Using Recon for Data Cleaning, In Proceedings of KDD.95: First hdernational Conference on Knowledge Discovery and Data Mining, pp. 275-281, Menlo Park, CA: AAAI Press.

Smyth, P., Burl, M., Fayyad, U. and Perona,P. 1996.Modeling Subjective Uncertainty in Image Annotation, in AKDDM, AAAI/MIT Press, 517-5, 10.

Stolorz, P. et al. 1995. Fast Spatio-Temporal Data Mining of Large Geophysical Datasets, In Proceedings of KDD-95: First International Conference on Knowledge Discovery and Data Mining, pp. 300-305, AAAI Press.

Therling,K (1995) An Overview of Data Mining at Dun and Bradstreet. DIG White Paper.

Turban, E., Aronson, J. E., Liang, T. P., & Sharda, R. (2007). Decision support and business intelligence systems (8th ed.). Taiwan: Pearson Education.

Violino,B. (2004), "BI for the masses", Computerworld, Vol. 38 No. 25, pp. 38-9.

Wang, J. (Ed.) (2005), Encyclopedia of Data Warehousing and Mining, Idea Group Inc., Hershey, PA.

Weiss, S. H., & Indurkhya, N. (1998). Predictive Data Mining: A Practical Guide. San Francisco, CA: Morgan Kaufmann Publishers.