Evolution of Data Mining: An Overview

Majid Ramzan[1], Majid Ahmad[2]
[1,2]PY-King Saud University, Riyadh, Saudi Arabia
[1]bramzan.c@ksu.edu.sa, [2]mcharoo.c@ksu.edu.sa

***Abstract:** Knowledge has played a significant role in every sphere of human life. To acquire knowledge we have to analyze the unlimited data that is available to us in various formats in the form of databases. We can analyze this data and find hidden information with the support of data mining. Data mining refers to the process or method that extracts interesting knowledge from large amounts of data. Data mining have number of applications and these applications have enhanced the various fields of human life including business, education, social media medical, scientific etc. The field of data mining has seen enormous success from the inception, in terms of wide-ranging application achievements and in terms of scientific advancement and understanding. The key objective of this paper is to provide an overview of evolution of data mining from its beginning to the present stage of development.*

***Keywords:** Data Mining; Data Mining Trends; Social Media;Big Data.*

I. INTRODUCTION

The 21st century known as information age, has affected every sphere of human life in the form of moderation of education (E- learning), medical, banking, sports, business etc. This resulted into large volumes of data stored in the form of numerical figures and text documents, to more composite information such as spatial data, multimedia data, and hypertext documents. To take complete advantage of data, the data retrieval is simply not enough, it requires a tool for automatic summarization of data, extraction of the information stored, and the discovery of patterns in raw data. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, to develop powerful tool for analysis and understanding of such data and for the extraction of interesting knowledge that could help in decision-making. The only solution to all above is 'Data Mining'. Data mining is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses [1, 3, 4]. Data mining tools predict future trends and behaviors, helps organizations to make proactive knowledge-driven decisions [2]. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer the questions that usually were time consuming to resolve. They prepare databases for finding hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Data mining, popularly known as Knowledge Discovery in Databases (KDD), it is the nontrivial extraction of implicit, previously unknown and potentially useful information from databases [3, 5]. Though, data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process [1, 3, 5].

II. DEVELOPMENTS IN DATA MINING

The term data mining was introduced in 1990s. Data mining touched its current state after going through numerous stages of study and research. This growth began when data started to get stored on computers. The process sustained with increase in computer capability including data storage, processing power, software etc. In today's world of technology, all are trying to make the optimal use of their data to make best decisions. Gathering and storing data on computers, tapes and disks started in 1960s.With the use of relational databases and structured query languages in 1980, helped users to do analysis about the data stored in relational databases using structured query language. Therefore, data became accessible at record level dynamically. In 1990 data warehousing was introduced. Multidimensional databases and online analytic processing contributed to the growth of data warehousing. Now, the emerging technique is data mining. If each step of evolution is studied, it is very clear that each step is constructed upon the preceding step.[11] To make key business decisions, managers need real time information. That information is provided by data mining techniques. During 1960s data was not considered as asset but the situation is now completely changed. Data has been changed to information which is sufficient to answer many questions and even to predict the future of business. Evolution of data and databases is happening at very fast which demand methods to deliver useful information from these large quantities of data. Data mining expertise have been going through growth process for many years and four different areas contributed to the growth of data mining in its current form. These areas are artificial intelligence, machine learning, statistics and databases. Statistics has been contributing significantly to business intelligence from the inception. The concepts of statistics deal with data and relations among them. These concepts are the building blocks of sophisticated data mining techniques. Artificial intelligence is the concept which is used to generate human thinking process or human intellect in statistical

978-1-4799-3064-7/14/$31.00©2014 IEEE

problems. Machine learning gives computers the capability to learn without being explicitly programmed. Database is the basic requirement for organized data mining. It is defined as collection of related data.

z

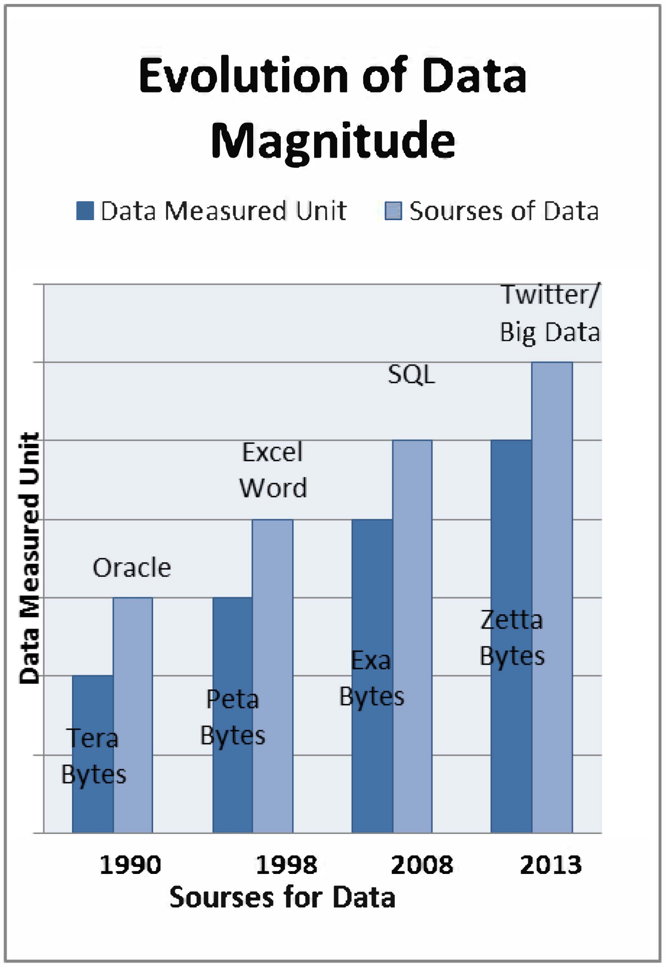

Fig. 1: Evolution of Data Magnitude

III. CURRENT TRENDS AND APPLICATIONS

A. Data Mining for Election Campaign

Data mining to target voters is new, of course. Before television, direct marketing and analytics software, campaign workers in local surroundings knew plenty about their neighbors, and used that information to make personal appeals. Later, candidates mashed up public records with consumer marketing data to develop advertising and fundraising appeals. Now, there's more data, it's centrally managed and may include people's social connections. Meanwhile, independent advocacy groups that support candidates for specific issues, but which aren't related to official campaigns or parties, also collect data from the same types of sources to target voters.

The Obama 2012 campaign used data analytics [big data[11]] and the experimental method to assemble a winning coalition vote by vote. In doing so, it overturned the long dominance of TV advertising in U.S politics and created something new in the world. A national campaign run like a local ward election, where the interests of individual voters were known and addressed. [28].

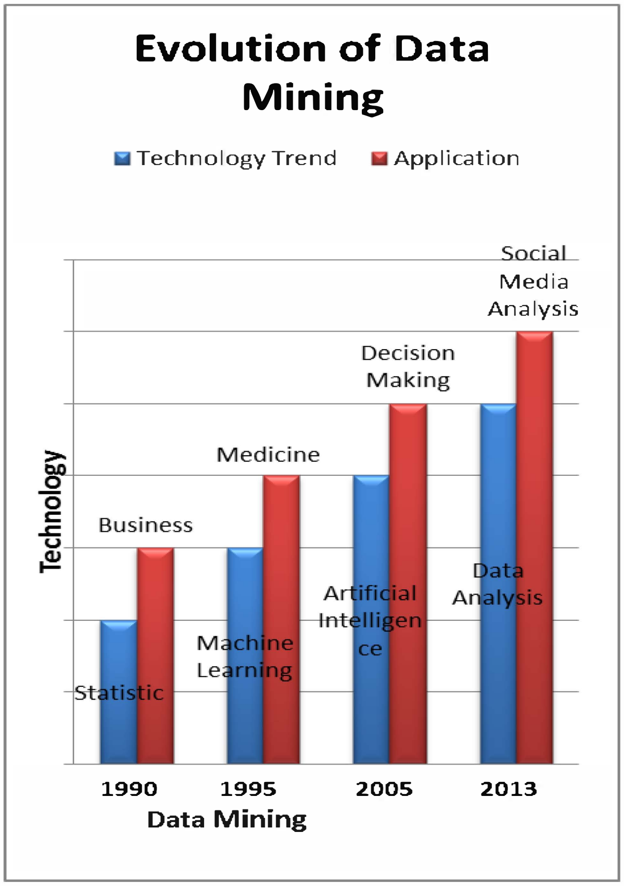

Fig. 2: Evolution of Data Mining

B. Data Mining for Social Media

Applying data mining methods to social media is relatively new compared to other areas of study related to social network analytics when you consider the work in social network analysis that dates back to the 1930s [26]. However, applications that apply data mining techniques developed by industry are already being used commercially. For example Facebook, twitter, Gmail provides services to mine and

[1] Big data is the term for the collection of data sets large and complex

monitor social media to provide organizations information about users. Researchers in other organizations have applied text mining algorithms to blogs to develop approaches for better understanding how information moves through the blogosphere [19].Data mining techniques can be applied to social media to understand data better and to make use of data for research and business purposes.

Representative areas include community or group detection [14, 24, 27], information diffusion [19], influence propagation [17, 12, 28, 25], topic detection and monitoring [19, 22], individual behavior analysis [15, 20, 21], group behavior analysis [23, 13], and of course, marketing research for businesses [10]. In the first half of 2013, Twitter made $32 million by selling its data namely tweets to other companies, a 53% increase from the year before. Because of its real-time nature, Twitter is the primary contributor to data mining, though other social networks are frequently used in professional analysis. [6]

C. Data Mining for Surveillance

It is a rising discipline, deals with developing methods for monitoring of a person or group's actions by organization. The data collected used for marketing purposes are sold to other organizations, but is also regularly shared with government agencies. It can be used as a form of business intelligence, which enables the organization to better improve their products and services to be eye-catching by their customers. The data can be sold to other organizations, so that they can use it for the decision making. It can also be used for direct marketing purposes, such as the advertisements on Google where ads are targeted to the user of the search engine by analyzing their search history and emails which is kept in a database. For example, Google, the world's popular search engine, stores identifying information for each web search. An URL address and the search phrase used are stored in a database for up to 18 months.

Google also scans the content of emails of users of its Gmail webmail service, in order to create targeted advertising based on what people are talking about in their personal email correspondences.[9]. Most of the companies monitor e-mail traffic of their workers, and 70% of corporations monitor Internet connections.

The United States government often gains access to these databases. The FBI, Department of Homeland Security, and other intelligence agencies have formed an information-sharing partnership with over 34,000 corporations as part of their Infrared program. The U.S. Federal government has gathered information from grocery store "discount card" programs, which track customers' shopping patterns and store them in databases, in order to look for "terrorists" by analyzing shoppers' buying patterns [9].

D. Web Content Mining

It aims to extract useful information from the content of a web page or website. It includes extraction of organized information from web pages, identification, match, and integration of semantically similar data, opinion extraction from online sources, and concept hierarchy, ontology, or knowledge integration [7]. Web content mining identifies the useful information from the Web Contents. However, such a data in its broader form has to be further narrowed down to useful information. The web content data consist of structured data such as data in the databases, unstructured data such as free texts, twitter, Facebook and YouTube and semi-structured data such as html, xml etc. Two main approaches are used in Web Content Mining: (1) Unstructured text mining approach and (2) Semi-Structured and Structured mining approach [8].

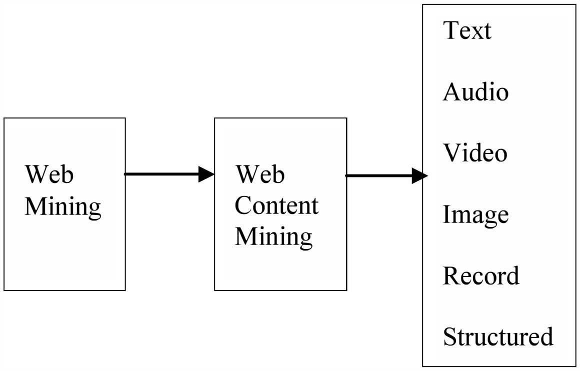

Fig. 3. Web Content Mining

IV. CONCULSION

In this paper we briefly reviewed the evolution and various data mining trends from its beginning. This review would be helpful to researchers to focus on the various issues of data mining. In future course, we will review the various big-data mining techniques and web mining techniques for customer reliability and satisfaction.

V. ACKNOWLEDGMENT

Majid Ahmad is a Scholar, working in the area of wireless Ad-hoc Networks. He is a member of IACSIT. He has almost five years of Research & Teaching experience in wireless communication, Information Storage management, and Software engineering.

Majid Ramzan is working as an IT Trainer, IT-Sills department at King Saud University Riyadh. He has an research interest in data-mining, Big Data and four years teaching experience in Computer science & information technology

REFERENCES

[1] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

[2] Larose, D. T., “Discovering Knowledge in Data: An Introduction to Data Mining”, ISBN 0-471-66657-2 John Wiley & Sons, Inc, 2005.

[3] Dunham, M. H., Sridhar S., “Data Mining: Introductory and Advanced Topics”, Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006.

[4] Chapman, P., Clinton, J., Kerber, R., Khabaza, T.,Reinartz, T., Shearer, C. and Wirth, R.. “CRISP-DM1.0 : Step-by-step data mining guide, NCR Systems Engineering Copenhagen (USA and Denmark),DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringenen Bank Group B.V (TheNetherlands), 2000”.

[5] Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., “From Data Mining to Knowledge Discovery inDatabases,” AI Magazine, American Association for Artificial Intelligence, 1996.

[6] www.natlawreview.com/article/twitter-s-data-mining-profits-show-lesser-known-social-media-risk

[7] Web Info Extractor Manual. WIhttp://webinfoextractor.com/wiedoc.html

[8] Johnson, F., Gupta, S.K., Web Content Minings Techniques: A Survey, International Journal of Computer Application.

[9] http://www.carefusion.com/medical-products/infection prevention/surveillance-analytics/medmined-data-mining-surveillance-service.aspx

[10] N. Agarwal and H. Liu. Modeling and Data Mining in Blogosphere, volume1 of Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan and Claypool, 2009.

[11] http://www.sqldatamining.com/index.php/data-mining-basics/history-of-data-mining

[12] N. Agarwal, H. Liu, S. Subramanya, J. Salerno, and P. Yu. Connecting sparsely distributed similar bloggers. pages 11 –20, Dec. 2009.

[13] P. K. Akshay Java and T. Oates. Modeling the spread of influence on the blogosphere. Technical Report UMBC TR-CS-06-03, University of Maryland Baltimore County, 1000 Hilltop Circle Baltimore, MD, USA March 2006.

[14] E.-A. Baatarjav, S. Phithakkitnukoon, and R. Dantu. Group recommendation system for facebook. pages 211–219, 2010.

[15] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography. In Proceedings of the 16th international conference on World WideWeb, pages 181–190, New York, NY, USA, 2007. ACM.

[16] Y. Chi, S. Zhu, K. Hino, Y. Gong, and Y. Zhang. iolap: A framework for analyzing the internet, social networks, and other networked data. Multimedia,IEEE Transactions on, 11(3):372 –382, april 2009.

[17] P. Domingos andM. Richardson. Mining the network value of customers. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pages 57–66, New York, NY, USA, 2001. ACM.

[18] C. Faloutsos, J. Han, and P. S. Yu., editors. Link Mining: Models, Algorithms and Applications. 2010.

[19] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In Proceedings of the 13th International Conferenceon World Wide Web, pages 491–501, New York, NY, USA, 2004. ACM.

[20] H. Lauw, J. C. Shafer, R. Agrawal, and A. Ntoulas. Homophily in the digital world: A livejournal case study. Internet Computing, IEEE, 14(2):15 –23, march-april 2010.

[21] Z. Liu and L. Liu. Complex network property analysis of knowledge cooperation networks. pages 544 –547, may 2009.

[22] J. Ritterman, M. Osborne, and E. Klein. Using prediction markets and twitter to predict swine flu pandemic. In F. M. Carrero, J. M. Gomez, B. Monsalve, P. Puertas, and J. C. a. Cortizo, editors, Proceedings of the1st International Workshop on Mining Social Media, pages 9–17, 2009.

[23] L. Tang and H. Liu. Toward collective behavior prediction via social dimension extraction. Intelligent Systems, IEEE, PP(99):1 –1, 2010.

[24] L. Tang, H. Liu, J. Zhang, N. Agarwal, and J. J. Salerno. Topic taxonomy adaptation for group profiling. ACM Trans. Knowl. Discov. Data, 1(4):1–28, January 2008.

[25] B. Ulicny, M. Kokar, and C. Matheus. Metrics for monitoring a social political blogosphere: A malaysian case study. Internet Computing,IEEE, 14(2):34 –44, march-april 2010.

[26] S. Wasserman and K. Faust. Social Network Analysis: Methods and Applications. Cambridge University Press, 1994.

[27] D. Zhou, I. Councill, H. Zha, and C. Giles. Discovering temporal communities from social network documents. In Seventh IEEE InternationalConference on Data Mining, pages 745 –750, Oct. 2007.

[28] www.technologyreview.com/featuredstory/509026/how-obamas-team-used-big-data-to-rally-voters