© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 08 Issue: 06 | April - June 2017
Paper is available at www.jrps.in | Email : info@jrps.in

ROLE OF DATA MININING IN WEB INTELLIGENCE: A REVIEW

Sohrab Ansari [1], Humera Zabin [2], Mohd Iqbal Mir [3]
AL-FALAH UNIVERSITY
Al-Falah School of Engineering & Technology
Sohrab.ansari90@gmail.com[1], humerzabin21@gmail.com[2], Iqbalmir06@gmail.com [3]

Abstract: The process of knowledge discovery in databases, often also called data mining, is first important step in knowledge management technology. End users of these tools & systems are at all levels of management operative workers & managers. And these are their demands on processing & analysis of data & information that affect development of these tools. Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. The challenges consist of analysis, data curing, capture, search, storage, sharing, transfer, visualization, and querying and information privacy. Soft computing is use of inexact solutions to computationally hard tasks such as solution of NP-complete problems, for which there is no known algorithm that could compute an exact solution in polynomial time.

ISSN : 2278-6848 03
9 772278 684800
© International Journal for Research Publication and Seminar

Keywords—Data mining, web mining, web intelligence, knowledge discovery, fuzzy logic, clustering

[1] INTRODUCTION

Data mining the analysis step of "Knowledge Discovery in Databases" process, or KDD, an interdisciplinary subfield of computer science, is computational process of discovering patterns in large data sets involving methods at intersection of artificial intelligence, machine learning, statistics, & database systems. Overall goal of data mining process is to extract information from a data set & transform this into an understandable structure for further use. Aside from raw analysis step, this involves database & data management aspects, data pre-processing, model & inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, & online updating.

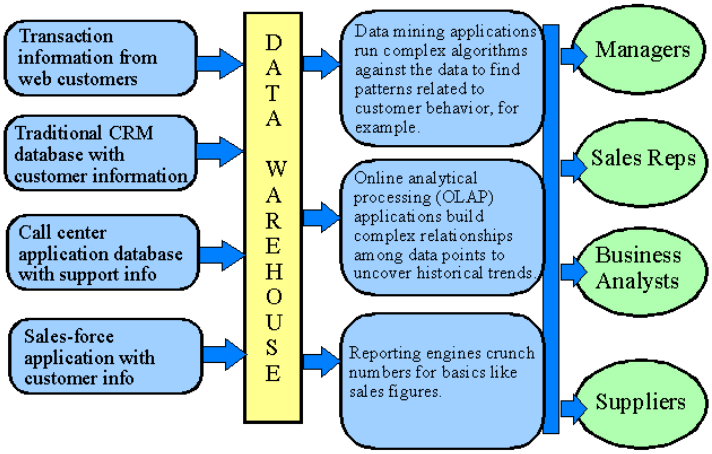

Fig 1 Data Mining process

The term is a misnomer, because goal is extraction of patterns & knowledge from large amount of data, not extraction of data itself. It also is a buzzword & is frequently applied to any form of large-scale data or information processing as well as any application of computer decision support system, including artificial intelligence, machine learning, & business intelligence. The famous book of "Data mining:

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 08 Issue: 06 | April - June 2017
Paper is available at www.jrps.in | Email : info@jrps.in

Practical machine of learning tools & techniques with Java" that covers the most of machine learning material has to be original named "Practical machine learning", & the term data mining has to be added for marketing reasons. Mostly the universal terms of data analysis, or analytics – or by referring to actual methods, machine learning & artificial intelligence – are most appropriate. The actual data mining task is automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records, unusual records & dependencies. This thing normally includes by using database techniques like spatial indices. These patterns could then be seen as a kind of summary of input data, & may be used in further analysis or, for example, in machine learning & predictive analytics

[2] MOTIVATION & PROBLEM STATEMENT

The Google Analytics which are Web Intelligence based are known as a service which was offered by Google. It generates some detailed statistics more about a website's traffic & the traffic sources & which measures conversions & sales. It's most widely used website statistics service. Basic service is free of charge & a premium version is available for a fee. Google Analytics may track visitors from all referrers, including search engines & social networks, direct visits & referring sites. It also tracks display advertising, pay-per-click networks, email marketing, & digital collateral such as links within PDF documents. Google Analytics Integrated with Ad Words allows users to review online campaigns by tracking landing page quality & conversions (goals). Goals may include lead generation, sales, viewing a particular page or downloading a specific file. Google Analytics approach is to show high-level, dashboard -type data/information for casual user, & more in-depth data/information further into report set. Analysis of Google Analytics might be identified by poorly performing pages by techniques like funnel visualization, the place where visitors came from the referrers, how long time they stayed there & about their geographical position. It provides enhanced features which includes the custom visitor segmentation. Google Analytics e-commerce reporting may track sales activity & performance. e-commerce reports show a site's transactions, revenue, & many other commerce-related metrics. Dashboards give you a summary of many reports on a single page. Start with a dashboard with your most important performance indicators (your "Company KPIs"), then create detailed dashboards for other important topics like search engine optimization. Dashboards use drag-and-drop widgets for fast, easy customization. Challenging problem in Web Intelligence is how to deal with uncertainty of information on wired & wireless Web.Adapting existing soft computing solutions, when appropriate for WI applications, must incorporate a robust notion of learning that will scale to Web, adapt to individual user requirements, & personalize interfaces. Ongoing efforts exist to integrate logic, artificial neural networks, probabilistic & statistical reasoning, fuzzy sets, rough sets, granular computing, genetic algorithm, & other methodologies in soft computing paradigm, to construct a hybrid approach/system for Web intelligence. Web intelligence techniques/technology for this level include Web data/information prefacing systems built upon Web surfing patterns to resolve issue of Web latency. The Intelligence of Web perfecting derived from adaptive learning process that is based on the observation & characterization of the behaviour of user surfing . Web may be considered as an interface for interaction of human-Internet. WI techniques/technology for this level are used to develop intelligent Web interfaces in which capabilities of adaptive cross-language processing, personalized multimedia representation & multimodal data/information processing are

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 08 Issue: 06 | April - June 2017
Paper is available at www.jrps.in | Email : info@jrps.in

required. Web is regarded as a distributed data/knowledge base. We need to develop semantic mark-up languages to represent semantic contents of Web available in machine-understandable formats for agent-based autonomic computing, such as searching, aggregation, classification, filtering, managing, mining, & discovery on Web.

[3] SURVEY OF EARLIER WORK

The use of data mining techniques in manufacturing began in 1990s & this has gradually progressed by receiving attention from production community. These techniques are now used in many different areas in manufacturing engineering to extract knowledge for use in predictive maintenance, fault detection, design, production, quality assurance, scheduling, & decision support systems. Data could be analyzed to identify hidden patterns in parameters that control manufacturing processes or to determine & improve quality of products. A major advantage of data mining is that required data for analysis could be collected during normal operations of manufacturing process being studied & this is therefore generally not necessary to introduce dedicated processes for data collection. Since importance of data mining in manufacturing has clearly increased over last 20 years, this is now appropriate to critically review its history & application. Data mining techniques becomes basic element of modern business. Although idea is not new, new technologies & implemented standards make a contribution to their growing popularity. Regarding to mining model usage SQL Server 2005 stands breakthrough in this area. Thanks to DMX language either programmers or database administrators are able to create Data Mining Systems in simple way. Although economical & business publications are very fruitful of data mining approaches, described problem is presented rather weak in international publications. Nethertheless some industrial appliances of data mining technology were considered in (Duebel, C., 2003). Industrial usage of data mining techniques opens new possibilities in decision making not only for top level management, but also for advisory or control systems. Several prediction, classification or even anomaly detection algorithms implementation may become lucrative tool for industrial process appropriate stages optimization, that combines diagnosis & control functions. The reviewed literature shows that there is a rapid growth in application of data mining in industry & manufacturing. However, there is still slow adoption of this technology in some industries for several reasons including both difficulties in determining type of data mining function to be performed in any particular knowledge area & question ofchoice most appropriate data mining technique regarding to many possibilities. **Waldemar Wójcik & Konrad Gromaszek (*Lublin University of Technology,Poland) introduced* "Data Mining Industrial Applications".** Data mining is blend of concepts & algorithms from machine learning, statistics, artificial intelligence, & data management. With emergence of data mining, researchers & practitioners began applying this technology on data from different areas such as banking, finance, retail, marketing, insurance, fraud detection, science, engineering, etc., to discover any hidden relationships or patterns.

Jiawei Han & Jing Gao University of Illinois at Urbana-Champaign wrote paper on "Research Challenges for Data Mining in Science & Engineering"

With rapid development of computer & information technology in last several decades, an enormous amount of data in science & engineering has been & will continuously be generated in massive scale, either being stored in gigantic storage devices or °owing into & out of system in form of data streams. Moreover, such data has been made widely available, e.g., via Internet. Such tremendous amount of data, in order of tera- to peta-bytes, has

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 08 Issue: 06 | April - June 2017
Paper is available at www.jrps.in | Email : info@jrps.in

fundamentally changed science & engineering, transforming many disciplines from data-poor to increasingly data-rich, & calling for new, data-intensive methods to conduct research in science & engineering. In this paper, they discuss research challenges in science & engineering, from data mining perspective,

[4]TOOLS & TECHNOLOGY USED

HARDWARE

- CPU 1Ghz or more
- HARDDISK (5GB Free space)
- DVD ROM
- MONITOR
- KEYBOARD/MOUSE

SOFTWARE

- WINDOWS 7/8
- MATLAB
- DOT NET FRAMEWORK

K-means clustering is a well known partitioning method. In this objects are classified as belonging to one of K-groups. result of partitioning method is a set of K clusters, each object of data set belonging to one cluster. In each cluster there may be a centroid or a cluster representative. In case where we consider real -valued data, arithmetic mean of attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases. K-Means Clustering algorithm is an idea, in which there is need to classify given data set into K clusters; value of K (Number of clusters) is defined by user which is fixed. In this first centroid of each cluster is selected for clustering & then according to chosen centriod, data points having minimum distance from given cluster, is assigned to that particular cluster. Euclidean Distance is used for calculating distance of data point from particular centroid. This algorithm consists of four steps: **Initialization:** In this first step data set, number of clusters & centroid that we defined for each cluster. **Classification:** distance is calculated for each data point from centroid & data point having minimum distance from centriod of a cluster is assigned to that particular cluster.

Centroid Recalculation: Clusters generated previously, centriod is again repeatly calculated means recalculation of centriod.

Convergence Condition: Some convergence conditions are given as below: Stopping when reaching a given or defined number of iterations. Stopping when there is no exchange of data points between clusters. If all of above conditions are not satisfied, then go to step 2 & whole process repeat again, until given conditions are not satisfied.

[5]CONCLUSION

The Internet of Things concept arises from need to manage, automate, & explore all devices, instruments, & sensors in world. In order to make wise decisions both for people & for things in IoT, data mining technologies are integrated with IoT technologies for decision making support & system optimization. Data mining involves discovering novel, interesting, & potentially useful patterns from data & applying algorithms to extraction of hidden information Due to increasing amount of data available online, World Wide Web has becoming one of most valuable resources for information retrievals & knowledge discoveries. Web mining technologies are right solutions for knowledge discovery on Web. Knowledge extracted from Web could be used to raise performances for Web information retrievals, question answering, & Web based data warehousing. The overall goal of data mining process is to extract information from a data set & transform this into an understandable structure. In data mining K-means clustering algorithm is one of efficient unsupervised learning

algorithms to solve well-known clustering problems. disadvantage in k-means algorithm is that, accuracy & efficiency is varied with choice of initial clustering centers on choosing this randomly.

References

1. J. Liu, S. Zhang, Y. Ye, Agent-based characterization of web regularities, in N. Zhong, et al. (eds.), Web Intelligence, NewYork: Springer, 2003, pp. 19–36.

2. J. Liu, N. Zhong, Y. Y. Yao, Z. W. Ras, wisdom web: new challenges for web intelligence (WI), J. Intell. Inform. Sys.,20(1): 5–9, 2003.

3. Congiusta, A. Pugliese, D. Talia, & P. Trunfio, Designing GridServices for distributed knowledge discovery, Web Intell. Agent Sys, 1(2): 91–104, 2003.

4. J. A. Hendler & E. A. Feigenbaum, Knowledge is power: semantic web vision, in N. Zhong, et al. (eds.), Web Intelligence: Research & Development, LNAI 2198, Springer, 2001, 18–29.

5. N. Zhong & J. Liu (eds.), Intelligent Technologies for Information Analysis, New York: Springer, 2004.

6. Bouckaert, Remco R.; Frank, Eibe; Hall, Mark A.; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter; Witten, Ian H. (2010). "WEKA Experiences with a Java open-source project".

7. Journal of Machine Learning Research **11**: 2533–2541. original title, "Practical machine learning", was changed ... term "data mining" was [added] primarily for marketing reasons.

8. Mena, Jesús (2011). Machine Learning Forensics for Law Enforcement, Security, & Intelligence. Boca Raton, FL: CRC Press (Taylor & Francis Group). ISBN 978-1-4398-6069-4.

9. Piatetsky-Shapiro, Gregory; Parker, Gary (2011). "Lesson: Data Mining, & Knowledge Discovery: An Introduction". Introduction to Data Mining. KD Nuggets. Retrieved 30 August 2012.

10. Kantardzic, Mehmed (2003). Data Mining: Concepts, Models, Methods, & Algorithms. John Wiley & Sons. ISBN 0-471-22852-4. OCLC 50055336.

11. "Microsoft Academic Search: Top conferences in data mining". Microsoft Academic Search.

12. "Google Scholar: Top publications - Data Mining & Analysis". Google Scholar.

13. Proceedings, International Conferences on Knowledge Discovery & Data Mining, ACM, New York.

14. SIGKDD Explorations, ACM, New York.