Journal of Advances and Scholarly Researches in Allied Education
Vol. 15, Issue No. 11, November-2018, ISSN 2230-7540

Data Mining: A Study on Data Cleaning Issues

Ravikant*

VPO-Munda, DISTT- Hanumangarh, Rajasthan

Abstract – Data quality is a big problem in quality information management. Data quality create problems in information system. These issues are resolved by data cleaning. Data cleaning is a procedure used to determine incorrect, incomplete or unhelpful data and then upgrade the quality through correcting of make out errors and exceptions. Correcting errors and eliminating redundant records can be time taken and unexpected process but it cannot ignored. Data mining is a methods for locating useful information in data. Data quality mining is a current approach used data mining methods to identify and remove data quality issues in large databases. Data mining automatically express hidden and inherent information from the group of data. Data mining has various methods that are fitting for data cleaning.

Key Words – Data Mining, Data Cleaning Tasks and Issues.

- X -

INTRODUCTION

Data cleaning, also named like called data cleansing or scrubbing, deals with discovering and eliminating errors and unpredictability from data in order to increase the quality of data.

Data quality issues are present in single data groups, like files and databases, e.g., misspellings entered during data entry, uncompleted information or other invalid information.

When more data origin need to integrated, e.g., in data warehouses, confederated database systems or web-based information systems, the need for data cleaning upgrade importantly. This is because the origin usually contain duplicate data in different presentations. In order to allocate access to correct and consistent data, consolidation of dissimilar data characterization and elimination of redundant information become useful.[1,2,3]

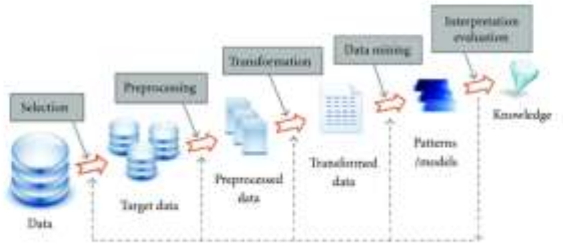
DESCRIPTION OF BROAD AREA

Data Mining

Data mining concern to bring out or mining knowledge from large group of data. Thus, data mining should have more accurately named as knowledge mining which significance on mining from large group of data.

It is the computational procedure of extracting patterns in large group of data sets involving techniques at the intersection of artificial intelligence, database systems, machine learning, and statistics.

The overall aim of the data mining procedure is to discover information from a group of data and convert it into an understandable data structure for future use.[4,5,6]



***Data mining as a step in Knowledge Discovery*[7,8]**

The key properties of data mining are

- Automatic discovery of patterns.
- Prediction of likely outcomes.
- Creation of actionable information.
- Focus on large databases.[9]

Tasks of Data Mining

The data mining tasks categories the common properties of data job execute inference on the present data set to rectify new data set behaviour.

There are several data mining tasks. Each task solved by a data mining algorithm because task can be considered as a kind of problem.[10,11,12]

Anomaly detection – The characterization of useless data sets, that might be attractive or data errors that require further investigation.

Association rule – Searches for associations (alliances) between variables. For example, a flipkart or amazon company might gather data on customer purchasing or searching habits. Using some association rule learning, those companies can determine which products have more frequency bought together and use this information for marketing benefits. Sometimes it is also called, market basket analysis.

Clustering – Is the process of finding structures and sets or data points into number of groups in the data that are in another "similar", without using any information about structures in the data.

Classification – Is the process of generalizing known structure to put in to new data. For example, e-mail program might make an effort to categories an e-mail as "legitimate" or as "spam".

Regression – Try to find a function, which models the data with the least error.

Summarization – Providing a more small scale rendering of the data set, together with report creation and visualization.

Data Mining - Issues

Data mining is very difficult task because always data is not present on single place, it need to be confederate from different sources and mining algorithm is be complex. These factors may create different issues.[13,14,15,16,17,18]

Performance of Algorithm:-Algorithm is essential part of data mining. The performance of data mining is depend on algorithms used. If the mining algorithm is not exquisite then result can be unexpected and it can be effect final data.

Mining dissimilar types of knowledge in databases. - The need of different users can be different and Different user may be in absorbed in different kind of knowledge. Therefore, it is required for data mining to embed broad range of knowledge discovery process.

Incorporation of unobtrusive knowledge. - Predictive tasks can become actual predictions and descriptive tasks can produce results that are more accurate so background knowledge can be used. Background knowledge may be exploit to demonstrate the recognize patterns not only in brief terms but also at multiple level of abstraction. Collecting and implementing background knowledge is a time taking and difficult process for data mining organizations.

***Protection and Privacy of Data* -** Privacy and protection of data is biggest issue for all the government and private organizations. The information reorganized by data mining can be valuable, people fear about another side of the coin, specifically the privacy threats by data mining. Individual's privacy may be infringe due to the unauthorized admittance to personal data.

Evaluation of Pattern - The patterns reorganized should be absorbing because either they present ordinary absence of originality or knowledge.

***Restructuring of Complex Data* -** Data that exist in real world also in different forms. Data can be in numeric form, audio form, graphical form, video form, text form, etc. collecting the data from different forms and computing the required information from different mediums of data can be complex.

Handling incomplete or noisy Data -The data in real world is unstructured and noisy. The data is in large quantity and will be unreliable due to the human error or by instruments. The data cleaning technique are necessary that can tackle the noise, unfinished entity while mining the data regularities. If data cleaning techniques are not there then the correctness of the recognize patterns will be low paid.

Data mining result's Visualization - First of all user able to capture properties of the data mining result that why it is important. Once the patterns are recognize it required to be represented in high-level languages. That representation need to be easily and understand by the users. After that visualization technique for some data mining methods like decision trees, association rules, clustering, and Bayesian networks.

Distributed, Parallel and incremental mining algorithms. - The elements those who massive size of databases, broad division of data and complication of data mining techniques influence the evolution of analogous and distributed data mining algorithms. These algorithms split the data into partitions, which is additionally process analogous. Then the outcome from the partitions is integrated. The upgrading algorithms, without mining the data again from scratch, update the database.

CONCLUSION

Data mining has enhance one of the key features of security of nation and grooming of business.

Often used as a means for detecting fraud, discover previously unknown, relationships in large data sets, Data mining helps businesses understand which marketing campaigns will likely generate the most engagement, classify customers, display personalized advertisements, and optimize marketing spend. To extract the patterns and the knowledge from variety of databases data mining uses the different methods and the Selection of data and methods for data mining.

Currently in the type of scientific tools, data mining is in significant advance but it has also imperfections. One imperfection is that however data mining don't notify the value or substance of the patterns. A second imperfection is that while data mining can recognizes relationship between variables, it don't automatically recognizes a causal connection. Successful data mining still have necessary trained technical and scientific professional who can formation the evaluation and explicate the outcome.

We often analyze customer data for understanding individual behavior and based on this understanding. Researchers in demographics, economics, medicine and social sciences are trying to find the relationships between behaviors and results. Finally, for the transparency of the costs and consequences to consumers, businesses, and the economy of legislative or regulatory proposals to protect privacy and security.

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. (2005). "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, ohn Wiley & Sons, Inc.

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

[4] Larose, D. T. (2005). "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, ohn Wiley & Sons, Inc.

[5] AC Yeo, KA Smith, RJ Willis and M Brooks (2002). Journal of the operation research society: 2002, A mathematical programming approach to optimize insurance premium pricing within a data mining framework.

[6] Han J. et. Kamber M. (2002). "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, Canada.

[7] "Discovering Knowledge: Introduction of Data Mining", ISBN 0-471-66657-2, ohn Wiley & Sons, Inc.

[8] "Discovering Knowledge in Data: An Introduction of Data Mining", ISBN 0-471-66657-2, ohn Wiley & Sons, Inc.

[9] Evfimievski, A., R. Srikant, R. Agrawal, and J. Gehrke (2002). "Privacy Preserving Mining of Association Rules," Proceedings of Eighth International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada, July 2002

[10] R. Agrawal, T. Imielinski and A. Swami (1993). Mining association rules in large databases. In: P. Buneman, S. Jajodia (Eds.) International Conference on Management of Data. ACM Press.

[11] E. Backer (1995). Computer-Assisted Reasoning in Cluster Analysis. Prentice-Hall.

[12] U. M. Fayyad and K.B. Irani (1993). Multi-interval discretization of continuous-valued attributes for classification learning. *(IJCAI '83)*.

[13] M. S. Chen, J. Han, and P. S. Yu (1996). Data mining: An overview from a database perspective. IEEE Trans. Knowledge and Data Engineering.

[14] Ding, P. (2009). A framework for Data Mining models. In Computational Intelligence and Industrial Applications.

[15] H. Vernon Leighton and J. Srivastava (1997). Precision Among WWW Search Services (Search Engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos. http://www.winona.msus.edu/isf/libraryf/webind2/webind2.htm.

[16] Cooley, B. Mobasher and J. Srivsatava (1997). Web Mining: Information and Pattern Discoveryon the Word Wide Web. In 9th IEEE International Conference on Tools with AI (ICTAI,97).

[17] R. Agrawal, T. Imielinski, and A. Swami (1993). Mining association rules in large databases. SIGMOD'93.

[18] R. J. Bayardo (1998). Efficiently mining long patterns from databases. SIGMOD'98.

Corresponding Author

Ravikant*

VPO-Munda, DISTT- Hanumangarh, Rajasthan

ravikantdaal92@gmail.com