Knowledge Discovery in Database Systems

Naciye Güliz Uğur and Aykut Hamit Turan

***Abstract* - Advancements in information and communication technology have propelled both the burgeoning of information appliances such as sensors, infrared, and smartphone/wear devices to capture data, as well as the ability to store and process such data, such as Hadoop and MapReduce. Consequently now, nearly all facets of industry and study are accompanied by informatics and computational resources [15]. This environment characterizes a data-intensive paradigm marked by intensive computation.**

Data in itself, though, especially in large volumes, is not helpful to decision makers unless converted to information, which is the collection of the relevant and helpful set of facts revealed by the data. When enough information about an attribute, topic or industry is acquired, it can be assumed that the individual is knowledgeable, and can make good decisions. Informatics is the field of gleaning knowledge from data—specifically, how to acquire and structure data, extract information from that data, and ultimately transform the information into useful knowledge for the domain at hand.

This study evaluates and discusses knowledge discovery trends and serves some valuable insights regarding business informatics and analytics, big data analytics and data mining techniques.

***Keywords* – Big Data, Data Mining, Data Science, Informatics**

I. INTRODUCTION

Advancements in information and communication technology have propelled both the burgeoning of information appliances such as sensors, infrared, and smartphone/wear devices to capture data, as well as the ability to store and process such data, such as Hadoop and MapReduce. Consequently now, nearly all facets of industry and study are accompanied by informatics and computational resources [1]. This environment characterizes what Jim Gray called the fourth paradigm—a data-intensive paradigm marked by intensive computation [1].

Data in itself, though, especially in large volumes, is not helpful to decision makers unless converted to information, which is the collection of relevant and helpful, action-ready, set of facts revealed by the data. When we have acquired enough information about an attribute, topic or industry, we can say that we are knowledgeable, and can make good decisions. Informatics is the field of gleaning knowledge from data—specifically, how to acquire and structure data, extract information from that data, and ultimately transform the information into useful knowledge for the domain at hand [1].

The particular leg of informatics concerned with extracting information from data is the domain of analytics. The face of analytics has changed drastically given the role of information and communication technology; notably, there has been a proliferation of information appliances to capture data such as smartphones, smart appliances, wearable technology, recording devices, infrared tolls, QR codes, registrations, check-ins, and social media applications; these appliances are capturing both structured and unstructured data; these various forms and sources of data can be delivered quickly through high-speed networks; and the data is both stored and computed upon at large scale data centers which have both the ability to store and compute the data.

This new makeup of high volume data from diverse sources requires a data-driven inquiry as Jim Gray predicted [1]. Data-driven is to be distinguished from the hypothesis-driven domain of statistics. The applicability of statistics, which has maintained a monopoly on data analysis for centuries, can be reexamined concerning large data: with the new quantity and makeup of data, is it feasible to proceed with an assumption driven, that is, “hypothesize-and-test” framework. Statistics are based on relatively smaller, cleaner, sampled data sets with a particular problem in mind, which poses a significant scalability issue, compounded with increasingly unstructured forms of data, which are difficult if not impossible for traditional statistical methods to process [2].

This new face of large, diverse, quickly acquired data is often referred to by its buzzword name “Big Data”, which already has been facetiously truncated to “data” given its ubiquity. Big Data, characterized by a high volume, variety, and velocity of data, is no longer an intellectual challenge for those in ivory towers, but is instead the daily reality for industry across the board, from infrastructure and healthcare to marketing and finance to cybersecurity and law; the volume of data has forced changes in traditional database storage, the velocity of data has forced computation to take place at the point of collection, and its variety has necessitated the fusion of data points which are not easily correlated [3].

Machine learning techniques have absorbed mainly the analytics evolution toward a data-driven inquiry [4], and this work will explore knowledge discovery trends and serves some valuable insights regarding informatics and analytics, big data analytics and machine learning techniques.

II. BIG DATA ANALYTICS

Much like statistics, data mining’s function is to input data and output knowledge [5]. However, data mining can better handle the salient characteristics of Big Data, which include high volume, velocity, and variety, and can better facilitate the increased computational intensiveness that Big Data demands.

While the term big data may appear to reference the volume of information, which is not generally the case. Big data, mainly when utilized by vendors, may allude to the innovation (which incorporates devices and procedures) that an association requires in managing large amounts of information and data. The term big data ought to have come into being with web search organizations that wanted to query massive accumulation of unstructured data.

Big data is transforming the way employees work within an enterprise. It is creating an atmosphere within which organizations and software professionals must work hand-in-hand to extract value from all sorts of data. Big data insights can help professionals in better decision making and taking advantage of newer revenue sources.

Machine learning techniques to mine data, both supervised and unsupervised, are illustrations of the data-driven approach - i.e., let the algorithms unveil the associations, classifications, and clusters from the data that can help predict and describe behavior [4].

Big Data analytics' usefulness lies within its predictive and descriptive powers; the underlying "facts" are the same as the days preceding Big Data. The three significant advantages of big data are:

Competitive advantage: Big data is evolving as the newest source for competitive advantage in the present world scenario.

Decision making: Big data is enabling more mid-level and lower-level employees to participate in the decision-making process of the organization.

Value of data: As the data becomes more valuable, there is a need for more sophisticated systems to extract meaningful information.

For instance, in 1974, if person B walked to the store to buy toilet paper and toothpaste, this fact may only have been known because person B told his spouse that he did so. Person B may have paid cash, did not pass by any security cameras, and did not have a GPS enabled device on his person. In 2017, it could be forensically verified from apartment building cameras that person B left his apartment at 4:16 pm; his smartphone could tell us he took exactly 1,740 steps to the store; social media confirming that he "checked-in" on Foursquare at 4:30pm to get his 15% coupon for the two items which he paid for on his Maestro card, whose database kept a record of this transaction date, time, amount, and location; all confirmed by the street camera and store camera that visually tracked his travel from apartment to store. While at its core, nothing that transpired is different - person B went to the pharmacy to buy toilet paper and toothpaste. So, the many statistics about the growth of data refer more to documentation production than in transpiring action: social media storing tweets that were once own private thoughts and conversations; servers storing emails and IMs that were once bar conversations and in-person meetings memorialized by notepads.

Of course, technology and consequently the computation on Big Data have injected new information at tremendous rates. Injecting new facts is not necessary to gain new information, but rather, it is the ability to record the facts or transpiring acts, which can then be analyzed to produce new, helpful descriptive and predictive information. The age of Big Data has allowed just that - it has allowed for the massive storage of digital information, the ability to compute at the point of collection, and incorporation of machine learning analysis to draw predictive and descriptive information.

As addressed earlier, Big Data is not "Big Statistics". While statistics is best suited toward smaller, cleaner datasets with an objective to understand why, Big Data, given its sheer volume, can permit imperfect, or less clean, data due to its size (with the caveat that there is not a systematic bias) [6], with more data points as input, more correlations and patterns will surface, whether or not anticipated or "relevant" to the inquirer. However, the lack of a pinpoint strategy or inquiry heading into the data is not a downfall of Big Data; instead, it has created a new framework for analysis using ad hoc inquiries that provide a strategy or course of action quickly for profit, health, or efficiency [6].

However, the age of Big Data has brought two salient challenges. First, Big Data can be too big. In statistics, sampling was critical because collecting data was expensive and perhaps challenging. In this age, the collection is easy. For many years, having more data for a decision maker was a dream come true.

However, as growth of data continued exponentially, the increased costs to store and compute began to take toll, and moreover, data mining and information fusion research results demonstrated that combining everything is not necessarily better, consistent with the fundamental statistical concept of regression toward the mean, developed well over a century ago by Sir Francis Galton [7]. Big Data analytics minimizes the amount of input needed for a more optimal result, and which can be applied at several levels as discussed in more detail further on.

A second characteristic distinguishing Big Data analytics from traditional statistics is that with more features requiring distinct methods for analysis, and perhaps producing distinct sets of data for prediction from another attribute and another method, correlation falls short - correlation involves dependencies between different data sets or different data distributions, and Big Data is producing independent viewpoints through its immense variety. Information fusion based technologies not only accommodate but thrives on, independent viewpoints.

III. Business Intelligence

BI is the process of gathering and analyzing information. The information must be relevant to the organization, regardless of the source. According to Spath [15], BI has become synonymous with information technology. However, when BI is used, it generates a vast amount of data to be converted into information and intelligence. Technology only acts like a tool. Ultimately, humans make all final decisions. Some strategies enhance BI, such as setting up systems and services designed to acquire, share, and disseminate information of all kinds, including customer encounters and market place data.

BI goes beyond technology with a focus on creating value and knowledge from information. When an effective knowledge management system is in progress, leverages can be created to add value to BI and vice versa. By adding a knowledge management system, the organization can attach meaning to information and effectively share information among staff members. Both BI and knowledge management can then be leveraged to create performance improvement through measuring, guiding, and motivating staff members.

Once data has been analyzed it becomes information to achieve improved organizational performance. Internal information flow is the process within an organization to control and manage knowledge sharing. Organizations that can remove barriers and allow for freer flowing information will benefit from significant cost savings. As similar to BI, knowledge management is not just information technology. Information technology is a tool that facilitates knowledge management. The goal of a knowledge management program is to deliver intellectual capacity to people who make day-to-day decisions [15]. Knowledge can reside in many different locations including databases, printed material, and people's head. High performing organizations have figured out how to share their knowledge across the entire system effectively.

IV. Data Mining

Data mining leads us to identify new, useful, and obvious correlations and patterns in existing data [16]. Different communities have different names for the same process of finding useful information (data) [17]. Statisticians, database researchers, and the MIS and Business communities started using the term "data mining" initially for extracting useful information from big data. One of the processes involving data mining is known as Knowledge Discovery in Databases (KDD), which is used for discovering useful knowledge from data. KDD involves data preparation, selection, cleaning and proper interpretation of results of the data mining process to ensure useful information is gleaned from the data. Data mining differs from traditional data analysis and statistical approaches in that it uses analytical techniques from several disciplines [17][18][19].

Several data mining tasks utilize a traditional, hypothesis-driven data analysis approach. It is also very commonplace to use an opportunistic, data-driven approach which aids pattern detection algorithm to find trends, patterns, and relationships which can then be used in the decision-making process. These two data mining approaches differ in the outcome – a model or a pattern. The model approach is similar to the conventional exploratory statistical methods, except for the problems inherent from the large size of data sets. The primary motive here is to summarize a set of data for identification and description of prominent features of distribution [20]. One of such models includes cluster analysis partition of a set of data, prediction using the regression model, and tree-based classification rule. While creating a model, sometimes empirical and mechanistic models are treated differentially [21]. Empirical (also called operational) models seek to establish relationships without any bias from any underlying theory. Mechanistic models (also called substantive or phenomenological models) are based on the theory of information producing period. Hence, data mining, by definition, is primarily concerned with operational.

Another type of data mining approach, pattern detection, seeks to identify small (and probably, significant) departures from the norm, which helps in detecting an unusual behavioral pattern, e.g., fraud detection for credit cards by monitoring unusual spending patterns, sporadic waveforms in EEG traces, etc. The notion that data mining seeks "nuggets" of information from a vast data repository, derived from this class of method. However, with the business database, it is not so easy to extract patterns, because of the complexity of data arising from anomalies as discontinuity, noise, ambiguity, and incompleteness [22]. The predictive power of such mining algorithms might decrease with an increase in some anomalies [23].

One of the most critical pre-processing steps in data mining is the construction of a data warehouse that involves data cleaning and integration. However, it is an optional step in most data mining process, as in case of more extensive data warehouse containing data from multiple sources – it becomes an enormous task, taking a tremendous amount of time running into years and costing millions of dollars [24]. Alternatives to data warehouses are an available operational or transactional database, or data marts, which can be either logical or a physical subset of the data warehouse.

Harmony of artificial intelligence and statistics-related technique leads to KDD systems, which aids in finding associations, sequences, classifications, clusters, and forecasts. Operational warehouse acts as the entry point for most of the data, which is then "cleaned" and moved into the warehouse. After a certain amount of time, these data are either purged or summarized (along with other information) or archived.

The term data mining is used to refer to a specific set of activity, which involve extracting meaningful new information from data. It is not a new term to statisticians and is synonymous with data dredging or data snooping in the hope of being able to find patterns. This happens if a dataset is re-used for inference or model choosing [25]. Accordingly, this "snooping" is a derogatory term for an exhaustive search can throw up a pattern of some kind while many of this pattern can be a product of random fluctuation and do not represent any underlying issue, which conflicts with the purpose of data analysis – to model the underlying structure giving rise to consistent and replicable patterns.

Hence, data mining is a tool that helps organizations to concentrate on the most significant data existing in their current database. This does not eliminate the need to know the business, the available data or the underlying analytical methods in use. Here, it must be noted, that the predictive method resulting from data mining may not be the cause of an action/behavior. Such causal inferences are subject to several error sources like inherent variability, sample selection bias and model equivalence of data population, or population drift [26]. Furthermore, it should also be noted, that relationship gleaned from the data mining process, does not indicate the value of pattern to an organization. Such a pattern should be verified and validated in an appropriate context.

Several industries benefit from using data mining and have witnessed increased profits by reducing costs and raising revenues. There are several ways to do this, a few of which are as follows: Helping in reducing costs during the beginning of the product life cycle in the research and development phase. Automated manufacturing processes use bounds for statistical processes, which can be derived from data mining - reducing mailing costs by avoiding mailing to customers who do not respond to offers.

Several enterprises take advantage of data mining to acquire new customers, increase revenue from existing customers, and to retain good customers, which is almost the whole customer life cycle. By using customer profiling, these companies can target prospective customers with similar traits and also focus attention on those customers who have not bought similar products (cross-selling).

While this sort of profiling enables a company to study the consumer behavior and attract new customers, it also enables them to retain customers who are considered at risk of leaving (called reducing churn or attrition). It is usually far less expensive to retain a customer than acquire a new one [27].

Data mining is such a ubiquitous subject that it can contribute to almost every stream of business. A few examples, where data mining can contribute are: Telecommunication and credit card companies - these are two of the leaders in the application of data mining to detect fraud. Insurance companies are yet another industry which uses data mining to reduce fraud. The medical industry uses data mining to predict the effectiveness of surgical procedures, medical tests, and medications. Financial firms use data mining to determine market and industry characteristics as well as to predict individual company and stock performance. Retailers can decide on products, which are popular, and items to stock in particular stores (even when and where to place them) and also access the effectiveness of promotions and coupons. Several Pharmaceutical firms mine large databases for chemical compounds and genetic materials to discover substances, which might be a potential candidate as new agents for the treatment of disease [29].

V. Machine Learning

Machine learning deals with a field of study, design, and development of different types of algorithms that provide the computers the ability to learn from the data without giving clear instructions. Machine learning algorithms can extract information automatically without any human help. Machine learning is a subfield of computer science and artificial intelligence. It also has a close connection with fields such as statistics and optimization. Kalman filter, optical character recognition, etc. can be termed as an example of a machine learning algorithm [28].

Machine learning algorithms can be organized into the following taxonomy based on the outcome that is desired:

Supervised Learning: These types of algorithm are the ones that are trained on examples called labeled examples where the inputs are provided with the desired output already known.

Unsupervised Learning: These types of algorithms are the ones that are trained on examples called unlabeled examples where the inputs are provided without the desired output being known.

Semi-supervised Learning: These types of algorithms operate on both labeled and unlabeled examples in order to produce the desired function.

Transduction: These types of algorithms try to generate new outputs that are based on fixed test cases from observed training cases.

Reinforcement Learning: Reinforcement learning is an area of machine learning that is concerned with the way software agents combine in an environment to maximize the reward or outcome.

Multi-tasked Learning: Multi-tasked learning is an area of machine learning that is trying to generate algorithms that learn a problem simultaneously with related problems taking into account a shared representation that leads to producing a better model for the initial task. This often results in a better model because it helps the learner to take out the best out of both the tasks involved.

Developmental Learning: Developmental learning also known as robot learning wherein the machine learns on its own based on human interactions and self-exploration and taking the help of guidance tools such as active learning, maturation, etc.

Machine learning and data mining can be confused, as they generally use the same techniques and overfit to a large extent. They can be roughly defined as follows: The difference between machine learning is that machine learning concentrates on predicting the outcome by learning from the data whereas in data mining the central aspect is to concentrate on finding the unknown traits in data set. It is also called the analysis step of knowledge discovery in databases. Both machine learning and data mining are closely correlated with data mining using many machine learning methods wherein the desired outcome is slightly different. Also, machine learning uses many data mining techniques such as unsupervised learning in order to improve learner accuracy [30].

As the diversity and quantity of data are growing each day, so too is the pool to comb for relevant information in litigation. Knowledge discovery's typical response to the explosion of electronically stored information has been to limit the amount of electronically stored information for human review through a keyword restriction. Aside from the fact that keywords have inherent semantic weaknesses that are highly vulnerable in Information Retrieval, data volume is growing so high that humans do not have the time to review the many false positives that result from such a method. In short, "linear review", even truncated by keyword restrictions, is becoming infeasible due to volume radically outpacing budgets and time, and new predictive coding techniques are raising the bar for acceptable, relevant production. The process now is fundamentally different – exponential growth needs to meet exponential review. Discovery must become an iterative, combined approach of humans training and supervising machine learning algorithms in order for attorneys to keep pace with data and to indeed certify they have produced everything relevant, as sworn to in the rules governing discovery. Incorporation of the combinatorial fusion paradigm into the human-driven part of the predictive coding cycle can even further enhance the quality of output.

VI. CONCLUSION

Information and Business Analytics can help companies, reduce costs while increasing quality which ensures competitive advantage in global or crowd markets. Information is an asset, and advancing organization performance requires organizations to manage this asset. Even with all the tools, data, analytics model, and data-driven reform organizations still struggle to identify and use the right data. There is little debate that big data and analytics is a must-have for supporting an organization's strategic goals.

Information systems and infrastructure cover the areas of computerized information systems and the establishment of a competent scalable infrastructure to enable strategic goals. High performing organizations have data systems that allow for universal access to information and enhanced decision making. Additionally, resources focusing on data activities should be consolidated under one umbrella with a multidisciplinary group setting expectations and providing direction.

The particular leg of informatics concerned with extracting information from data is the domain of analytics. The face of analytics has changed drastically given the role of information and communication technology; notably, there has been a proliferation of information appliances to capture data such as smartphones, smart appliances, wearable technology, recording devices, infrared tolls, QR codes, registrations, check-ins, and social media applications; these appliances are capturing both structured and unstructured data; these various forms and sources of data can be delivered quickly through high-speed networks; and the data is both stored and computed upon at large scale data centers which have both the ability to store and compute the data.

Successful information management refers to the choice of right tool or technique and the quality and integrity of data. Accuracy and timeliness of data are significant factors when making decisions. An overarching big data governing body can help establish and focus efforts on the validity of data created. Information analysis refers to the process of analyzing data to create value and knowledge. As previously stated, data does not equate to value. Data must be analyzed to create value. High performing organizations take efforts to invest in activities to create value from data. Lastly, the internal information flow refers to the transfer of information throughout the organization promptly. Emphasis on internal information flow requires that communication barriers must be broken down and allow for complete knowledge sharing.

Current work has shown that information analytics promises predictive and descriptive powers for companies and decision makers. The most critical problem to address in analyzing big data is the choice of a tool or technique to be used. It mainly depends on the jurisdiction of the analyst or the researcher. Therefore, the researchers have come up with four different techniques to analyze the given traffic data set. The goal is to help the analysts in understanding the different knowledge discovery techniques in order to find the best-fit solution for their particular problems.

Of course, technology and consequently the computation on Big Data have injected new information at tremendous rates. Injecting new facts is not necessary to gain new information, but rather, it is the ability to record the facts or transpiring acts, which can then be analyzed to produce new, helpful descriptive and predictive information. The age of Big Data has allowed just that - it has allowed for the massive storage of digital information, the ability to compute at the point of collection, and incorporation of machine learning analysis to draw predictive and descriptive information.

In summary, information analytics is a relatively new effort in businesses, and there is no clear universal path for success. Organizations should start to build systems (technology and resources) that include understanding all data sources and how to create value from data. The process to build these systems cannot be implemented overnight. Finding the right tools and techniques will provide a path and direction for organizations to create systems that are scalable and beneficial.

REFERENCES

[1] Hey, T., Tansley, S., and Tolle, K. (Eds.). Jim Gray on e-Science: A Transformed Scientific Method. Microsoft Research. 2009.

[2] Friedman, J. Data Mining and Statistics: What's the Connection? Department of Statistics and Stanford Linear Accelerator Center. 1998.

[3] Liu, H., Wu, Z. H., Zhang, X., and Hsu, D. F.: A skeleton pruning algorithm based on information fusion. Pattern Recognition Letters 34(10), (2013); p. 1138-1145.

[4] Mitchell, T. Machine Learning, WCB McGraw-Hill, 1997.

[5] Tan, P., Steinback, M., and Kumar, V. Introduction to Data Mining, Addison Wesley, 2006.

[6] Mayer-Schönberger , V. and Cukier, K. Big Data: A Revolution That Will Transform How We Live, Work and Think. Houghton Mifflin Harcourt. March 2014.

[7] Galton, F. Natural Inheritance. Long MacMillan & Co. 1889.

[8] Page, S. "Diversity and Complexity" Princeton university press. 2010.

[9] Deng, Y. , Wu, Z., Chu, C.H., Zhang, Q., Hsu, D.F.: Sensor Feature Selection and Combination for Stress Identification Using Combinatorial Fusion. International Journal of Advanced Robotic Systems 10, (2013); p. 1-10.

[10] Chung, Y. S., Hsu, D. F., Tang, C. Y. "On the diversity-performance relationship for majority voting in classifier ensembles", Multiple Classifier Systems, Lecture Notes in Computer Science, (2007) 4472:407-420.

[11] Chung, Y. S., Hsu, D. F., Liu, C. Y., Tang, C. Y.: Performance evaluation of classifier ensembles in terms of diversity and performance of individual systems. International Journal of Pervasive Computing and Communications. 6(4), (2010); p. 373 – 403.

[12] Zhou, Z. Ensemble Methods: Foundations and Algorithms; Chapman & Hall/CRC, 2012.

[13] Hsu, D. F., Chung, Y. S. and Kristal, B. S.; Combinatorial fusion analysis: methods and practice of combining multiple scoring systems, in: H. H. Hsu (Ed.), Advanced Data Mining Technologies in Bioinformatics, Odeal Group, (2006), pp. 32-62.

[14] Hsu, D. F., Kristal, B. S., Schweikert, C. Rank-Score Characteristics (RSC) Function and Cognitive Diversity. Brain Informatics 2010, Lecture Notes In Artificial Intelligence, (2010), pp. 42-54.

[15] Spath, P. L. Leading your healthcare organization to excellence. Chicago: Health Administration Press, 2005.

[16] H. M. Chung and P. Gray, “Special Section: Data Mining”, Journal of Management Information Systems, vol. 16, pp. 11-17, 1999.

[17] U. Fayyad et al., “From data mining to knowledge discovery in databases”, AI Magazine, vol. 17, pp. 37-54, 1996.

[18] P. R. Peacock, “Data mining in marketing: Part 1”, Marketing Management, pp. 9-18, 1998.

[19] Lico, L. Data Mining Techniques in Database Systems. Proceeding presented at International Conference on Sustainable Development. October, 2016. Skopje.

[20] Lico, L. Data Mining Techniques in Database Systems. Proceeding presented at International Conference on Sustainable Development. October, 2016. Skopje.

[21] D. J. Hand, “Data mining: statistics and more?”, The American Statistician, vol. 52, pp. 112-118, May 1998.

[22] G. E. P. Box and W. G. HUNTER, “Sequential design of experiments for non linear models”, in Proc. IBM Scientific Computing Symposium on Statistics, New York, pp. 113-137, 1965.

[23] U. Fayyad et al., "The KDD process for extracting useful knowledge from volumes of data," Communications of the ACM, vol. 39, pp. 27-34, 1996.

[24] B. Rajagopalan and R. Krovi, “Benchmarking data mining algorithms”, Journal of Database Management, vol. 13, pp. 25-36, Jan-Mar, 2002.

[25] P. Gray and H. J. Watson, “Professional Briefings...Present and future directions in data warehousing”, Database for Adv. in Info. Sys., vol. 29, pp. 83-90, 1998.

[26] H. White, “A Reality Check for Data Snooping”, Econometrica, vol. 68, pp. 1097-1126, 2000.

[27] C. Glymour et al., “Statistical Inference and Data Mining”. Communications of the ACM, vol. 39, pp. 35-41, 1996.

[28] M. J. Berry and G. S. Linoff, “Mastering Data Mining: The art and science of customer relationship management”. Wiley Computer Publishing, New York, 2000.

[29] Wikipedia, (2017). “Machine Learning” [Online]. Available: https://en.wikipedia.org/wiki/Machine_learning

[30] Kumar, V., Kumar, N. AND Pandey, K. M. Data Mining and Business Intelligence: Concept and Component. IJARSE, 4 (5), p. 188-191.

[31] J. Han and M. Kamber, “Data mining: concepts and techniques”, Morgan-Kaufmann Academic Press, San Francisco, 2001.