A COMPREHENSIVE SURVEY ON DATA MINING

Kautkar Rohit A[1]

[1]*M. Tech Scholar, Computer science and technology, Maharashtra Institute of Technology (MIT), Aurangabad, Maharashtra, India*

Abstract

Now a day's internet is a significant place for interchanging of data like text, images, audio, and video and for share-out information preferably in digital form. The usage of internet leads to accessing the immense amount of data. Data may be unstructured data, structure data, and semi-structured data. So we are storing and processing such vast amount of data having gigantic complexity. Researchers from the University of Berkeley estimate that every year about one Exabyte (= 1 Million Terabyte) of data brought forth, of which a large portion is in digital form.

For ex., 100 hours of videos are uploaded on YouTube every minute (statistics from YouTube). The query is how to analyze such ample data effectively and efficiently? Answer is data mining. Data mining pertains to the process of analyzing, studying such declamatory quantity of data for witnessing useful patterns and knowledge. Today we have accumulation of large amount of data but deficiency of knowledge.

In this paper our focusing on surveillance of a nimble arising field data mining which is also known as knowledge discovery from data (KDD). We are constituting fundamentals of data mining, also several strategies for analyzing data like classification, estimation, prediction, association rules, clustering etc., data mining process, its types like web, content and structure mining. After understating basics, we are presenting the different data mining models like decision tree, neural network with their bedrocks. Also presents real world applications and future scope of the dynamic and extraordinary discipline.

Keywords: *- Internet, Data, Unstructured Data, Structure Data, Semi-Structured Data, Data Mining, Knowledge Discovery from Data (KDD)*

----------***----------

1. INTRODUCTION

Now day's internet is a crucial place for interchanging the information and for sharing, dealing out the digital contents like text, audio, video, images etc. In that social networking sites acts crucial role like face book (www.facebook.com), LinkedIn micro blogging sites like Twitter (www.twitter.com) for sharing digital contents. So they produce, process huge and gigantic quantity of data every day in fact every minute.

For example According to public statistics more than 48 hours of video contents are uploaded every minute and billions views are rendered every day on YouTube (www.youtube.com). Researchers from the University of Berkeley estimate that every year about 1 Exabyte (= 1 Million Terabyte) of data are generated, of which a large portion is available in digital form [3]. For goes on updates it connects to social networking sites like Face book, Twitter etc. so amount of data produce is quite vast and may have size in millions megabytes (or more) and complex in structure. So it leads to use highly efficient, advanced tools and techniques for intension of analyzing, processing such data. Analyzing and processing of data allow comprehending useful information and knowledge about the data. The term "Data mining" is acquainted in 1990s [12]. So inquisitory for knowledge in data is nothing but data mining [13]. Mining is important since it grant learning about versatile trends lives in data [18].

For example, Google search engine which accept and process huge amount of information every minute. If it just have a large accumulation of data and do not have knowledge regarding it? Then it's pathetic. On other hand, if we employ data mining tools and techniques then we may find practicable patterns and trends. Google demonstrates different current trends according to geographical area which is example of gaining knowledge from user submitted queries.

Search engines and social networking sites like Google and face book produce, store and process huge amount of data in terms of volume, velocity and variety denoted as big data. Big data integrates structured, semi-structured and structured data. So in reality data mining go through big data which in huge in size and complex in structure. The ambitious matter is that, it is continuously and rapidly increasing.

There are several reasons for increasing the data on internet in rapid fashion. The use of internet leads to access, process and creation of data. For example, the use of face book leads sharing or uploading any images, videos. It's nothing but creation of data. Another example is use of twitter. When we post any twit then it contributes to creation of Data. The Company's which deals with huge amount of data everyday for them the field of data mining is vital.

The data over internet may available in different types. Figure 1 depicted its types like structure, semi-structure and structure. When we have vast amount of data then it produces big data which creates another challenge in every aspect. The sources of data may be internal or external. Such data analyzed by predefined tools and technique referred as data mining. It allows discovering utile trends and patterns in data.

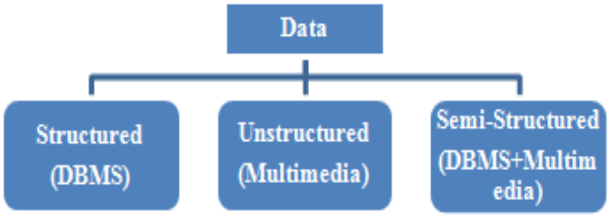

Fig 1: Types of Data

The simplest example of pattern, milk can be purchased with bread or computer can be purchased with Antivirus software [13].

According to Gartner group (www.gartner.com), data mining is the process of discovering meaningful correlations, patterns and trends by shifting large amount of data store in repositories, using pattern recognition technologies as well as statistical and mathematical techniques [12]. We have large amount of data available but don't have knowledge concerning it. So the data mining imparts a way for experiencing knowledge from data. For example, Wall mart a megastore, which records millions of transaction every day, give rise to creation of huge data [13]. So it we don't have a satisfying way to canvass it then its utility will be atrophied. So the usefulness can be enhanced by analyzing it with different tools and techniques.

At present, the data mining established its grandness in field of information retrieval and processing and there is no uncertainty that in near future, it will keep arising. Data mining also called Knowledge discovery in Database (KDD) i.e. the process of exploring useful patterns and knowledge from data. Knowledge discovery is the treatment of analyzing data for future projection [21]. To gratify the demand of analyzing the available large volume data sets, more efficiently the data mining employs diverse fields like Machine learning, Information retrieval, Statistics, and visualization.

Figure 2 shows; there are two primal intentions of data mining process- prediction and description. Prediction permit to influence the stranger or future values of pursuit where as description grant noticing patterns and describing the data [6]. The Meta Group estimates that the market size for data mining market will grow from $50 million in 1996 to $800 million by 2000 [11].

As shown in figure 3, the consolidation of diverse fields leads to new dynamic, encouraged area data mining. The data mining construct is actually formulated from statistics and machine learning. It also consolidates artificial intelligence, Information retrieval, visualization etc. We can conflate data mining with machine learning techniques for deceitful usage of cellular telephones grounded on profiling customers [29]. Fundamentally machines ascertain from data collected in past like humans learns from past experience and every time uses yesteryear experience to perform more beneficial. It is similar to supervised learning also referred as classification. It uses classification algorithms like decision tree or naive bayes [14]. When genuine utile pattern detected, then its usefulness can be determined with assist of statistics. We may visualize our data that to analyzed, the data visualization is ministrant for discovering interesting patterns easily. Basically computers cater easy mode to view data in more powerful and efficient manner. Now there are more powerful visualization tools are available [18]. The visualization tools can help for discovering approximate idea about the kind of data and its quality, also to anticipate where the useful patterns will situate [30].

At start-of paper, we stated, the internet is place for interchanging the digital contents over web. So it having gigantic size of information available and to explore accurate information in which user is interested is challenge. The search is most prominent practical application of web. The field which is participating in this problem is Text mining. The target of Information retrieval system is to elicit documents pertained to user query (a set of keywords) from huge amount of available information. The user's information demand can be provide in form of query to Search engines [14]. Text analysis is performed for ordering documents. After exploring documents, their degree of matching keyword query is used to rank those [19]. Different commercial algorithms used by search engines. Basically Information Retrieval (IR) Systems are employs Information agents [30] which search for appropriate documents. The challenge is retrieve relevant documents to user query with fast response. Also the web pages are not just text; they also with hyperlinks and anchored text [14].
Data mining system can be very complex or simple as it integrates different arenas. All subfields are important in data mining as they grant constructing solution to a greater extent complex problem. It also support for miscellany of data mining system [13]. So data mining system can be class based on measures like kind of database used for mining, kinds of knowledge mined and levels of abstraction of knowledge mined. Data mining system can be constructed based on what type of pattern we are exploring, regular or irregular [13]. Statistics are useful for assess data. The data may continuous or discrete. So different themes used to measure continuous data like average, mean, median, mode. To measure discrete data, histograms can be used which basically represent single values [31].

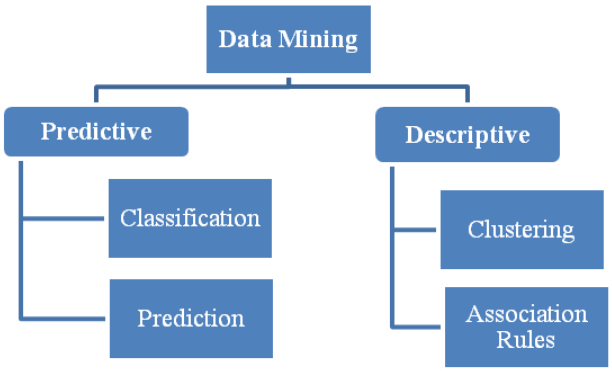

Fig 2: Data Mining Process Goal

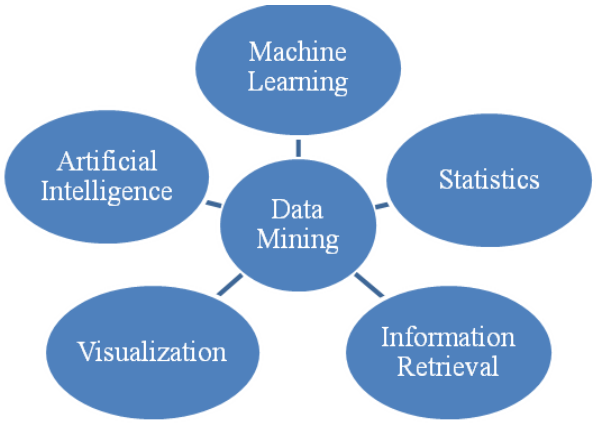

Fig 3: Association of data mining with other fields

2. RELATED WORK

Various data mining techniques that allow extracting unidentified relationships among the data items from large data collection that are useful for decision making [24]. Techniques for performing clustering on supervised data proposed [23]. Data mining techniques can play an important role in rule refinement even if the sample size is limited [8].

Artificial neural network models are proposed for imitate the human brain model. Decision tree allow segmenting the data in to appropriate groups. Both artificial neural network and decision tree have idea where to look for patterns in data. The use of market basket analysis & clustering techniques does not require any knowledge about relationships in the data, knowledge is discovered when these techniques are applied to the data. Market basket analysis tools sift through data to let retailers know what products are being purchased together. Clusters prove to be most useful when they are integrated into a marketing strategy [8]. Different marketing schemes shows that the firms profit can be increased by focusing on different marketing resources [25]. Apriori Algorithm [26] for data mining for frequent pattern mining proposed which uses prior knowledge about data. Different approaches to automatic construction of web pages by using log file data proposed [27]. The expected location of web pages can be handle page caching by browser. For this, algorithm automatically discovers pages in a website whose location is different from where visitors expect to find them [28].

The concept of a KDDM process model was primitively devised in 1989. Initial KD systems provided only a single DM technique, such as a decision tree or clustering algorithm, with a very imperfect support for the overall process framework. Initially it was using exclusive Data mining technique like association rule or clustering or decision tree.

In Advances Knowledge Discovery and Data Mining define the process as being highly interactive and complex [24]. They also hint that KDDM process may comprise aggregation of Data mining techniques for processing and observing more complex patterns [10].

3. DATA MINING STRATEGIES AND PROCESS

3.1 Data Mining Strategies

There are following strategies for analyzing data items-

1. **Classification** in which items are ordered /classified in to usable target classes.
2. **Estimation** allows ascertaining unidentified output variables.
3. **Prediction** allows determining the future consequence. It is similar to classification and estimation.
4. **Association Rules** allow analyzing the big data sets for detecting useful patterns and relationships between data items. Masses of applications are consume association rule concept.
5. **Clustering** allow grouping of items according to alike properties and behaviors. There are various algorithms are available for clustering like K-means.

3.2. Data Mining Process

Exploring utile patterns and knowledge from available data invokes fuse of steps as depicted in figure 4, like [13]-

1. **Data Cleaning**

As we go through, data is collected from internal or external sources and may contain noise and inconsistent information. Data cleaning phase allow getting rid of noise and inconsistent data. When data set is large, then data cleaning is time demanding process.

2. **Data integration**

The data may available in scattered structure. So data integration phase allow to blend it from different sources. Phase 1 and 2 are considered as preprocessing phases and accompanying data may be stored in to data warehouse. (Not mandatory).

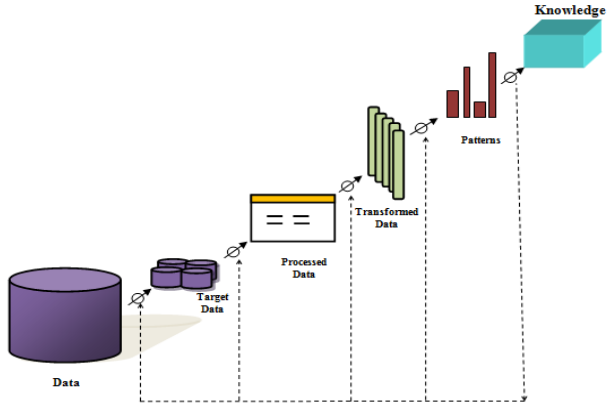

Fig 4: Data Mining Process [22]

3. Data selection

Allow retrieving data from database for analyzing. According to problem domain, the different data sources can be selected.

4. Data transformation

The data transformed into the form which will be appropriate for processing and analyzing.

5. Data mining

Actual grasping of useful knowledge and patterns in data is achieved by data mining.

6. Pattern evaluation

In this, noticing truly interesting patterns, various standards like lift, support, confidence etc. can be used.

7. Knowledge presentation

Visualizing knowledge and representation technique can be used to constitute results.

4. DATA MINING MODELS

There are many different data models are available for figuring out the data mining problems. The models are like decision tree, neural network, naive bayes, classifiers, lazy learners etc.

4.1 Decision Trees

It is also mentioned as classification tree. The decision tree allows distinguishing data and classifying it into different subdivisions or groups.

As shown in Figure 5, it follows the divide and conquers scheme. As it is decision tree it permits drawing decisions by trialing different potential considerations. The node of tree depicting the testing of attributes values. It allows testing of condition with several relational and comparison operators. Decision tree can be used for both description and prediction. Various algorithms are available like ID5, C4.5 and C5 etc.

Decision tree act upon data that can be presents using table where each row is keyed out as instance shown in Table 1. Various useful patterns can be extracted from table for learning and knowledge intention as shown in figure 5.

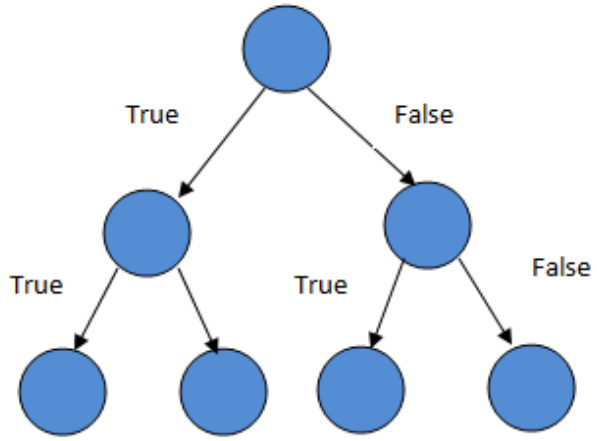

Fig 5: Decision Tree

Decision trees are interesting as they are based on rules. They are used to explore relationships amongst the data items. The decision trees are structure used to divide large data set in to small sets based on rules. Rules can be written into English language [31].

Table 1: Table with Different Instance of Data

Sr.No	Employee Name	Age	Gender	Salary	Item Buy
1	Ram	30	M	40000	TV
2	Chetan	35	M	52000	{TV ->VCD Player}
3	Prasad	25	M	25000	{Laptop->Antivirus Software}
4	Sangeeta	45	F	45000	{Laptop->Antivirus software}

4.2 Neural Network

The concept of neural network is fundamentally inherited from the construction of human brain. Basically the neural networks are used for acquiring knowledge from various data sets and amend the performance of system. In Neural Network the knowledge is presented using interconnected processors addressed neurons. Each neuron consist weight value and with defined function. [1, 5]

As indicated in figure 6, there are different input vectors(x_1, x_2…..x_m) with consorted weights (w_1, w_2…..w_m) and possible outcome positive or negative.

The neurons are cultivated to reacts each input vector. The vectors from training set is applied over Neural Network, if the output is right then no alteration is made otherwise the weights and bias are adjusted.

When the NN is fully trained, then it will classify the applied input vector by using its knowledge. Different classifiers can be used to classify the output.

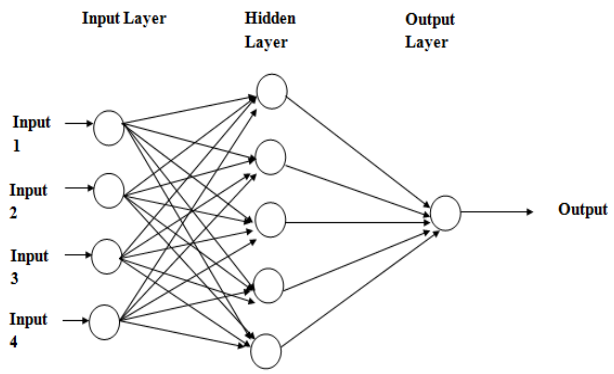

Fig 6: Neural Network

4.3 Association Rules

Association rules basically give the relation between different attributes and their values. Gobs of useful associations can be explored by analyzing the data sets. Association rule utilizes the concept of support and confidence to determine the usefulness of rule. Patterns are practicable for probing the buying habits of customer for business analysis.

Support for items X1 and Y1 for particular transaction can be given by percentage of transactions from data set which let in both items i.e. X1 U Y1. Confidence for association rule X1->Y1 is given by percentage of transaction from data set containing X1 also with Y1 [13]. The rule is genuinely useful if it has support and confidence value above the defined support count for that dataset. The support count can fix by domain expert [13].

4.4 Naive Bayes

The naive bayes constructs model that predict the probability of particular outcome.

The possible outcomes can be predicted by finding patterns and correlations between the data instances. Afterward it picks out the data mining model for representation of the patterns and relations.

4.5 Machine Learning and Statistics

As we know the data mining field extensively uses the concepts of machine learning and statistics for mining complex patterns. The statistics and machine learning are really much pertained arenas and in much sense depend on each other. The statistics are basically used for constructing the hypothesis where the hypothesis is used in the machine learning for generalization process.

5. TYPES OF MINING

5.1 Web Mining

The web mining grants eliciting information from web documents and services using data mining techniques. Web mining term devised by Etzioni [4].

Internet comprises immense volume of data and is the most declamatory data source for acquiring and apportioning information. The web mining pertains to the discovering useful information from the web hyperlink structure, web page contents and web usage data. There are three classes as shown in figure 7 [4]-

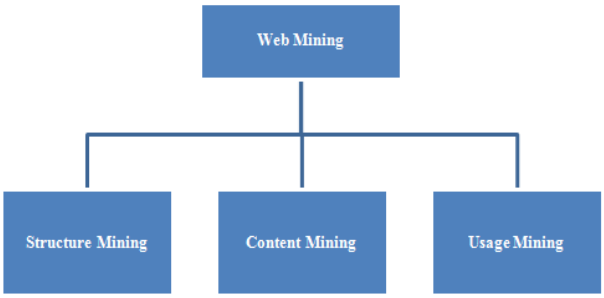

Fig 7: Web Mining Types

5.1.1 Web Structure Mining

Web structure mining has to do with discovering the useful knowledge is from the hyperlinks. Basically, hyperlinks represent structure of web. Here, we are interested in structure of web documents and also in structure of hyperlinks [4].

It as well applies concept of social network analysis in which the graph can be used to represent the structure. For example, on twitter, the graph based database (FLOCK DB) can be utilized to represent users with their followers.

5.1.2 Web Content Mining

Basically it grants to extract the useful patterns and knowledge from the web page contents. It is based on traditional mining process which allows classifying and clustering web pages (referred as items) according to their types of contents [4].There are two different standpoints are available in concern with web content mining-

1) Information Retrieval 2) DB View

Basically information retrieval is for semi structured and unstructured data. It permits querying the data and finding useful information. Database view is basically for the structured information having proper organization for storage and queries.

5.1.3 Web Usage Mining

Web usage mining rivets the technique that could predict user behavior whenever the user interacts with web. The data is viewed in terms of interactivity with web application. The web usage mining basically uses the web logs and server logs for analyzing it. The usage data can be depicted with relational tables or graphs. The usage mining is applies concepts like association rules, machine learning etc. for mining aim. The usage mining is important at many places like web application construction, marketing etc.

The web usage mining is useful in learning the user profile for providing better personalized experience. It is also helpful in the dynamic recommendation systems. Web usage mining permits to study user access patterns for giving better experience [19]. The issue of privacy arises; it exposes the privacy because the user profile data is used for clustering and analyzing the user profiles.

6. APPLICATIONS OF DATA MINING

The data mining is applicable in panoptic and diverse areas. The data mining tools and techniques are really play important role in fields where data is integral, highly sensitive and important [3].

1. Retail Industry

Huge amount of data is generated and stored. Data may comprise transaction details of sales, customer shopping etc. The stored data can be utilized for ascertaining the buying habits of customers, ameliorating the business, supervising the shelf space etc.

2. Telecommunication Industry

In telecommunication industry it can be useful for improving service quality, making better use of resources etc.

3. Biological Data Analysis

The data mining is useful here, to interpret the biological sequence and structure.

4. Semantic web

The web pages are analyzed and organized according to their types of contents. It utilizes the proficiency referred as Resource description framework (RDF) for classifying web pages. The RDF is implemented at lots of websites like orkut, face book for tagging purpose.

5. Business Trends

To serve customer more accurately, quickly and efficiently the data mining can be used.

6. Financial Data Analysis

Data collected from different sources of financial sectors and analyzed using data mining.

7. Sports

Many games played worldwide. So on each day many games are planned and played. It causes the vast amount of data creation. The data about each game and player can be analyzed and used to predict the performance of player.

8. Manufacturing Process

In manufacturing the data mining is useful for finding faults in manufacturing process and those faults can be amended. The data is collected from the manufacturing system.

7. CONCLUSIONS

This study paper canvassed the extremely dynamic and substantial area data mining. In this paper we confronted, the basics of data mining, why data mining is important, different strategies for data mining like classification, estimation, prediction, clustering, and association rules. Also we studied that many applications now day's based on strategies association rule for mining various trends. Also data mining process describes different phases in order to find the useful patterns and knowledge. It too introduces various data mining models like decision tree which allow making decision, NN which mimic the human brain structure and used for mining patterns more accurately, association rules for exploring relationships between data items. Naive bayes, machine learning and statistics are original ancestral of data mining field. The important portion of paper gives the type of web mining like structure mining, usage mining and content mining with their description. At long last it presents the diverse areas uses the data mining.

We also resolve that analyzing the huge amount of data with prominent complexity is challenging chore and data mining supplies a direction to deal with such data efficiently and effectively. The data mining is one of the spectacular and challenging field today's date and no doubts that, hereafter it will play vital primal in many areas.

ACKNOWLEDGEMENTS

I am enormously grateful to Prof. Mrs. K.V. Bhosale, Prof. V. Kala, Prof. R.B. Mapari, and Prof. Sonavane of Computer science and technology Department, MIT, Aurangabad.

As well thankful to Prof. M. M. Goswami, Prof. P. Ghotkar, Prof. Mrs. R.D. Kalambe, Prof. Mrs. M.S. Arade, Prof. Miss. N. S. Nikale, Prof. Mrs. S. P. Dudhe of Government Polytechnic, Nashik.

I genuinely thankful to Prof. J. Gorane,Prof. N. A. Anwat who directed me in right way for penning above work.

REFERENCES

[1] Singh, Y., & Chauhan, A. S. (2009). Neural Networks In Data Mining. Journal Of Theoretical And Applied Information Technology, 5(6), Pages 36-42.

[2] Keim, D. A. (2002). Information Visualization and Visual Data Mining. Visualization and Computer Graphics, IEEE Transactions On, Volume 8(1), Pages 1-8.

[3] Silwattananusarn, T., & Tuamsuk, K. (2012). Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012. Arxiv Preprint Arxiv: 1210.2872.

[4] Kosala, R., & Blockeel, H. (2000). Web Mining Research: A Survey. ACM Sigkdd Explorations Newsletter, Volume 2(1), Pages 1-15.

[5] Lu, H., Setiono, R., & Liu, H. (1996). Effective Data Mining Using Neural Networks. Knowledge and Data Engineering, IEEE Transactions On, Volume 8(6), Pages 957-961.

[6] Chaudhary, R., Singh, P., & Mahajan, R. A SURVEY ON DATA MINING TECHNIQUES.

[7] Jain, N., & Srivastava, V. (2013). DATA MINING TECHNIQUES: A SURVEY PAPER. IJRET: International Journal of Research in Engineering and Technology, Volume 2(11).

[8] Hilage, T. A., & Kulkarni, R. V. (2012). Review of Literature on Data Mining. IJRRAS, Volume 10(1), Pages 107-114.
[9] Mann, A. K., & Kaur, N. (2013). Survey Paper on Clustering Techniques. International Journal of Science, Engineering and Technology Research, Volume 2(4), Pp-0803.
[10] Kurgan, L. A., & Musilek, P. (2006). A Survey Of Knowledge Discovery And Data Mining Process Models. The Knowledge Engineering Review, Volume 21(01), Pages 1-24.
[11] Goebel, M., & Gruenwald, L. (1999). A Survey of Data Mining and Knowledge Discovery Software Tools. ACM SIGKDD Explorations Newsletter, Volume 1(1), Pages 20-33.
[12] Sharma, M. Data Mining: A Literature Survey.
[13] Han, J., & Kamber, M. (2006). Data Mining, Southeast Asia Edition: Concepts and Techniques. Morgan Kaufmann.
[14] Liu, B. (2007). Web Data Mining. Springer-Verlag Berlin Heidelberg.
[15] Bramer, M. (2013). Principles of Data Mining. Springer.
[16] Witten, I. H., & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.
[17] Wattenhofer, M., Wattenhofer, R., & Zhu, Z. (2012, June). The Youtube Social Network. In ICWSM.
[18] Hand, D. J., Mannila, H., & Smyth, P. (2001). Principles of Data Mining. MIT Press.
[19] Markov, Z., & Larose, D. T. (2007). Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage. John Wiley & Sons.
[20] Larose, D. T. (2006). Data Mining Methods & Models. John Wiley & Sons.
[21] Bhosle, K. (2010). KDS for Sericulture Cocoon Production. International Journal of Computer Applications, Volume 1(18), Pages 86-90.
[22] Raval, K. M. Data Mining Techniques. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2(10), Pages 439-442.
[23] Ahmed, S., Coenen, F., & Leng, P. (2006). Tree-based partitioning of date for association rule mining. Knowledge and information systems, Volume 10(3), Pages 315-331.
[24] Fayyad, U. M. (1996). Data mining and knowledge discovery: Making sense out of data. *IEEE Intelligent Systems*, Volume *11*(5), Pages 20-25.
[25] Peppers, D. (1993). The one to one future: Building relationship one customer at a time.
[26] Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Volume. 1215, Pages. 487-499).
[27] Perkowitz, M., & Etzioni, O. (1999, July). Adaptive web sites: Conceptual cluster mining. In *IJCAI* (Volume. 99, Pages. 264-269).
[28] Srikant, R., & Yang, Y. (2001, April). Mining web logs to improve website organization. In *Proceedings of the 10th international conference on World Wide Web* (Pages. 430-437). ACM.
[29] Fawcett, T., & Provost, F. J. (1996, August). Combining Data Mining and Machine Learning for Effective User Profiling. In *KDD* (Pages. 8-13).
[30] Sumathi, S., & Sivanandam, S. N. (2006). Introduction to data mining and its applications. Springer.
[31] Berry, M. J., & Linoff, G. (1997). Data mining techniques: for marketing, sales, and customer support. John Wiley & Sons, Inc.

BIOGRAPHIES

Kautkar Rohit A, Pursuing M. Tech From MIT, Aurangabad. Since last three years he is assisting as a lecturer in Computer Technology at Government Polytechnic, Nasik.

He has accomplished B. Tech (Computer science and Engineering) from SGGSIE&T, Nanded. He also completed Diploma in Information Technology from Government Polytechnic, Nasik. His domain of pursuit is Data Mining as well Big data, Natural Language Processing and web associated fields.