Rungta International Journal of Computer Science and Information Technology (RIJCSIT) Volume 2, Issue 1&2, 2017
ISSN(O): 2456–026X

A Review Paper on Data Mining Techniques

Kavita Yadav[1*], Om Prakash Dewangan[2]
[1]Dept. of Computer Science & Engineering, Rungta College of Engineering and Technology, Bhilai, Chhattisgarh, India. Email: Kavitayadav1026@gmail.com
[2]Dept. of Computer Science & Engineering, Rungta College of Engineering and Technology, Bhilai, Chhattisgarh, India. Email: Dewangan.omprakash@gmail.com
*Corresponding Author

Abstract: Data mining refers to extraction of information from huge chunks of the dataset. It's also called information mining. It is exercised in numerous fields like medicine, environment, education, market and business analysis, fraud detection, customer retention, crime, etc. In this research work data mining, text mining and web mining techniques used for data analysis and discovers patterns. This review paper covers clustering techniques (K-Means clustering technique; density based clustering and cosine similarity), web mining and text mining techniques. Clustering helps to put objects into the same group. Cosine similarity measure helps in finding similarity among different texts.

Keywords: Clustering, Cosine similarity, Data mining, Density based clustering, K-means clustering.

I. Introduction

Data mining is complete subfield of computer science. The goal of the data mining process is to mining information from a huge data set and transform it into an understandable. It is the scanning setup of the "knowledge discovery in database" process. The term is a misconception because the goal is the extraction of knowledge and pattern from large amount of data, not the extraction of data itself.

Data Mining is the computation process of discovering pattern in large data set involving methods at the intersection of artificial intelligence, machine learning, statistics and database system. The data mining is mostly used to very large scale of information processing (collection of data, extraction, statistics and warehousing analysis) as well as any application of computer decision support system, including artificial intelligence and machine learning.

The actual data mining task is the automatic analysis of large quantities of data to extract previously unknown, interesting pattern such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining). Data mining is process of detecting valid, new potentially useful and eventually understandable pattern in data

Selection of correct data mining is closely related to the initial data analysis. To meet started objective, it is possible to use several different methods. This depends on the structure of the data and the requirements of a particular method of data mining on the input [1].

A. Data Mining

One of the definitions of the concept of data mining: "data mining is the process of analyzing data from different perspective and their conversion into use full information from the mathematical and statistical point of view it comes to finding correlation, thus interrelationship or patterns in the data" [3]. There are also other definitions that described data mining, however they which largely depend on the purpose of uses of data mining methods. Data mining is the process divided into multiple steps which are shown on the Fig. 1.

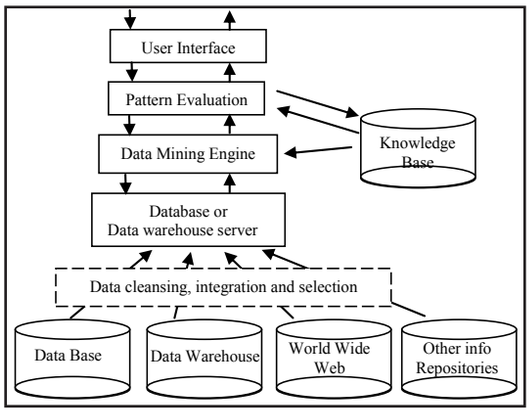


Fig. 1: Data Mining Architecture that Show Stages of Data Mining Process from Row Data to Knowledge

1. Major Data Mining Task
   - Classification: Predicting an item class.
   - Association: Examples A, B and C occur frequently.
   - Visualization: To facilitate human discovery.
   - Estimation: Predicting a continuous value.
   - Deviation Detection: Finding changes.
   - Link Analysis: Finding relationships.

2. Sources of Data Mining

Database (most obvious), text document, computer simulation and social networks.

3. Challenges of Data Mining
- Scalability
- Dimensionality
- Complex and heterogeneous data
- Data quality
- Data ownership and distributions
- Privacy presentations
- Steaming data

4. Advantages of Data Mining

Marketing, Retail, Finance, Banking, Manufacturing and Governments.

5. Uses of Data Mining
- AI/ Machine Learning
  Combinatorial /Game Data Mining
  Good for analyzing winning strategies to games, and thus developing intelligent AI opponents (i.e.; Chess)
- Business Strategies
  Marketing Basket Analysis
  Identify customer demographics, preferences and purchasing patterns.
- Risk Analysis
  Product Defect Analysis
  Analyze product defect rates for given and predict possible complications (read: lawsuits) down line.
- User Behaviors Validation
  Fraud Detection
  In the realm of cell phones
  Comparing phone activity to calling records. Can help detect call made on cloned phones.
  Can detect activity with stolen cards.
- Health and Science
  Protein folding
  Predicting protein interaction and functionality within biological cells.
  Applications of this research include determining causes and passible curse for Alzheimer's, Parkinson's, and some cancers (caused by protein "misfols")
  Extra-Terrestrials Intelligence
  Scanning Satellite receptions for possible transmission foremother planets.

B. Clustering

Clustering may be unsupervised classification of patterns like observation, data items, or feature vectors, into groups. The clustering problems can be found many contexts and is used by researchers in many disciplines.

Clusters are broadly classified into different categories bases on different criteria.

1. Representation based clustering or partitioning clustering - Partitioning Algorithms clustering algorithm splits the data points into k partition, where each partition represents cluster. The partition is done based on certain objective function. One such criterion functions is minimizing square error criterion which is computed as, $E = \Sigma \Sigma \| p - mi \|2$ Where p is the point in a cluster and mi is the mean of the cluster. The cluster should exhibit two properties, they are (a) each group must contain at least one object (b) each object must belong to exactly one group. The main drawback of this algorithm is whenever point is close to the center of another cluster; it gives poor result due to overlapping of data points [4]. It uses several greedy heuristics schemes of iterative optimization.
   There are many methods of partitioning clustering; they are k-mean, Bisecting K Means Method, PAM (Partitioning around Methods), CLARA (Clustering Large Applications) and the Probabilistic Clustering.

a) K-Means Clustering algorithm

K –Means clustering algorithm are given below
- Select K points as initial centroids.
- Repeats.
  - From K clusters by assigning each points to its closest centroid.
  - Precompute the criterion of each of cluster.
- Until the convergence criterion is met.

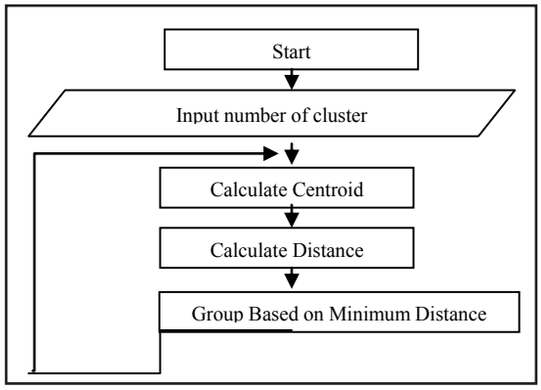


Fig. 2: K-Means Clustering Flow Chart

Mathematical example of K-Means Clustering:

Apply K-Means clustering for the following data set for two clusters. Tabulate all the assignments

Sample Number	X	Y
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77

Given K = 2 initial centroid

Cluster	X	Y
K1	185	72
K2	170	56

Calculate Euclidean distance using the given equation

Distance $E[(x,y),(a,b)] = \sqrt{(x-a)^2+(y-b)^2}$

Initial centroid

Cluster	X	Y
K_1	185	72
K_2	170	56

Cluster 1

$(185,72) = \sqrt{(185-185)^2+(72-72)^2} = 0$

Distance from cluster 2

$(170,56) = \sqrt{(170-185)^2+(56-72)^2}$

$= \sqrt{255+256}$

$= \sqrt{481} = 21.93$

Cluster 2 $(170,56) = \sqrt{(170-170)^2+(56-56)^2} = 0$

Cluster	Centroid		
	X	Y	Assignment
K_1	0	21.93	1
K_2	21.93	0	2

Calculate Euclidean distance for the next data set (168, 60)

Distance $E[(x,y),(a,b)] = \sqrt{(x-a)^2+(y-b)^2}$

Distance from cluster 1

$(185,72) = \sqrt{(168-185)^2+(60-72)^2}$

$= \sqrt{283+144}$

$= \sqrt{433} = 20.808$

Distance from cluster 2

$(170,56) = \sqrt{(168-170)^2+(60-56)^2}$

$= \sqrt{4+16}$

$= \sqrt{20} = 4.472$

Data Set	Euclidean Distance		
	Cluster 1	Cluster 2	Assignment
(168,60)	20.808	4.472	2

Update the cluster centroid

Cluster	X	Y
K_1	185	72
K_2	$(170+618)/2 = 169$	$(60+56)/2 = 58$

Calculate Euclidean distance for the next data set (179, 68)

Distance $[(x,y),(a,b)] = \sqrt{(x-a)^2+(y-b)^2}$

Distance from cluster 1

$(185,72) = \sqrt{(179-185)^2+(68-72)^2}$

$= \sqrt{36+16}$

$= \sqrt{52} = 7.211103$

Distance from cluster 2

$(169,58) = \sqrt{(179-169)^2+(68-58)^2}$

$= \sqrt{100+100}$

$= \sqrt{200} = 14.14214$

Data Set	Euclidean Distance		
	Cluster 1	Cluster 2	Assignment
(179,68)	7.211103	14.14214	1

Update the cluster centroid

Cluster	X	Y
K_1	$(185+179)/2 = 182$	$(72+68)/2 = 70$
K_2	169	58

Calculate Euclidean distance for the next data set (182, 72)

Distance $[(x,y),(a,b)] = \sqrt{(x-a)^2+(y-b)^2}$

Distance from cluster 1

$(182,70) = \sqrt{(182-182)^2+(70-70)^2}$

$= \sqrt{0+0}$

$= 0$

Distance from cluster 2

$(169,58) = \sqrt{(182-169)^2+(72-58)^2}$

$= \sqrt{169+196}$

$= \sqrt{365} = 19.10$

Data Set	Euclidean Distance		
	Cluster 1	Cluster 2	Assignment
(182,72)	2	19.10	1

Update the cluster centroid

Cluster	X	Y
K_1	$(182+182)/2 = 182$	$(70+72)/2 = 71$
K_2	169	58

Calculate Euclidean distance for the next data set (188,77)

Distance $[(x,y),(a,b)] = \sqrt{(x-a)^2+(y-b)^2}$

Distance from cluster 1

$(185,72) = \sqrt{(188-182)^2+(77-71)^2}$

$= \sqrt{36+36}$

$= \sqrt{72} = 8.4852$

Distance from cluster 2

$$(169,58) = \sqrt{(182-169)^2 + (77-58)^2}$$

$$= \sqrt{361+361}$$

$$= \sqrt{722} = 26.87$$

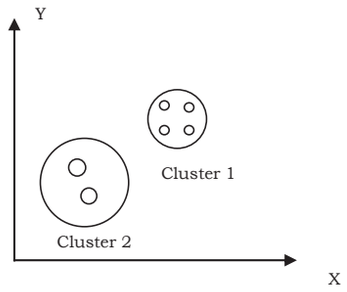
Data Set	Euclidean Distance		
	Cluster 1	Cluster 2	Assignment
(188,77)	8.485	26.87	1

Update the cluster centroid

Cluster	X	Y
K_1	$(182+182)/2 = 182$	$(71+77)/2 = 74$
K_2	169	58

Final assignment

Data Set	X	Y	Assignment
1	185	72	1
2	170	56	2
3	168	60	2
4	179	68	1
5	182	72	1
6	188	72	1



b) Kernel K-means

c) Expectation - Maximization clustering

This algorithm is applicable only when mean is defined (what about categorical data?). It requires specifying k, the number of clusters, in advance which is very difficult. It is not able to handle noisy data and outliers. It is not suitable to discover clusters with nonconvex shapes. For handling the categorical data, the algorithm k-modes are developed. It uses the latest dissimilarity procedures to deal with categorical objects and use a frequency-based method to revise modes of clusters. For a mixture of categorical and numerical data the k-prototype is used.

2. Hierarchical Clustering

Hierarchical clustering is a technique of clustering which divide the similar dataset by constructing a hierarchy of clusters. This method is based on the connectivity approach based clustering algorithms. It uses the distance matrix criteria for clustering the data. It constructs clusters step by step. Hierarchical clustering generally fall into two types: In hierarchical clustering, in single step, the data are not partitioned into a particular cluster. It takes a sequence of partitions, which may run from a single cluster containing all objects to 'n' clusters each containing a single object.

a) Agglomerative Nesting-

It is also known as AGNES. It is bottom-up approach. This method construct the tree of clusters i.e. Nodes. The criteria used in this method for clustering the data is min distance, max distance, avg distance center distance. The steps of this method are:

- Initially all the objects are clusters i.e. leaf.
- It recursively merges the nodes (clusters) that have the maximum similarity between them.
- At the end of the process all the nodes belong to the same cluster i.e. known as the root of the tree structure.

b) Devise Analysis-

It is also known as DIANA. It is top -down approach. It is introduced in Kaufmann and Rousseau (1990). It is the inverse of the agglomerative method. Starting from the root node (cluster) step by step each node forms the cluster (leaf) on its own. It is implemented in statistical analysis packages.

Advantages of hierarchical clustering

- Embedded flexibility with regard to the level of granularity.
- Ease of handling any forms of similarity or distance.
- Applicability to any attributes type.

Disadvantages of hierarchical clustering

- Vagueness of termination criteria.
- Most hierarchical algorithm does not revisit once constructed clusters with the purpose of improvement.

3. Density Based Clustering

Density based algorithms find the cluster according to the regions which grow with high density. It is the one-scan algorithms. It is able to find the arbitrary shaped clusters and handle noise. Representative algorithms include DBSCAN, GDBSCAN, OPTICS, and DBCLASD. The density based algorithm DBSCAN (Density Based Spatial Clustering of Applications with Noise) is commonly known. The Eps and the Mints are the two parameters of the DBSCAN. The basic idea of DBSCAN algorithm is that a neighborhood around appoint of a given radius (ε) must contain at least minimum number of points (Min Pts). The steps of this method are:

- Randomly select a point t.
- Recover all density-reachable points from t wrt Eps and MiPts.
- Cluster is created, if t is a core point.
- If it is a border point, no points are density-reachable from t and DBSCAN visits the next point of the Database.
- Continue the procedure until all of the points have been processed.

Density based algorithms classify into four types:

- The DBSCAN Algorithm

- Kernel Density Estimation
- Density Based Estimation and
- DENCLUE

4. Grid Density Based Algorithms

Grid Density based clustering is concerned with the value space that surrounds the data points not with data points. This algorithm uses the multiresolution grid data structure and use dense grids to form clusters.. Its main distinctiveness is the fastest processing time, since like data points will fall into similar cell and will be treated as a single point. It makes the algorithm autonomous of the number of data points in the original data set. Grid Density based algorithms require the users to specify a grid size or the density threshold, the problem here arise is that how to choose the grid size or density thresholds. To overcome this problem, a technique of adaptive grids are proposed that automatically determines the size of grids based on the data distribution and does not require the user to specify any parameter like grid size or the density threshold. The grid-based clustering algorithms are STING, Wave Cluster, and CLIQUE. These methods are efficient only for low dimensions. Among the huge number of cells most are empty and some may be failed with one point. It is impossible to determine the data distribution with such a coarse grid structure. Fine grid size leads to the huge amount of computation, while coarse grid size results the low quality of clusters. The algorithm OPTICS is proposed for the purpose of high dimensional data.

The steps of the grid based algorithm are:

- Making the grid structure, in other words divide the data space into a finite number of cells.
- Computing the cell density for each cell.
- Sorting of the cells according to their densities.
- Identifying cluster centers.
- Traversal of neighbor cells.

There are various algorithms used for clustering the data items into the clusters. Among them the Grid Density algorithms perform well over the time complexity as well as on the high dimensional data.

C. Cosine Similarity

In cosine similarity, words are used as a vector to find the normalized dot product of the two documents. The similarity scales between 0 and 1 (maximum). The bigger the return value is, the more similar the two text. Equation 1 shows the mathematical formula for finding similarity among two vectors.

$$Similarity\ (x,y) = cos\theta = \frac{x.y}{||x||\ ||x||\ ||x||\ ||x||}$$

$cos\theta$ = 1, it means that the two text are equal i.e. the documents share attributes

$cos\theta$ = 90, it means that the two text are totally different i.e. the documents share attributes

D. Web Mining

Web mining is the application of data mining techniques to extract pattern from the World Wide Web. Web mining gathered information by mining the. Web mining is the provision of information mining procedures to concentrate learning from Web information, i.e. Web Content, Web Structure and Web Usage information. WM is characterized as programmed creeping and extraction of applicable data from the relics, exercises, and concealed.

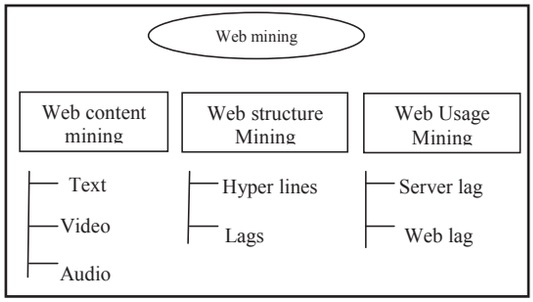


Fig. 3: Web Mining Techniques

Web mining divided into three different types-

1) Web Content Mining- Web content mining is the extraction and integration of data. Web mining terminology are:

   a) Web content mining has two types of view of data- IR View: Unstructured and Structured view of data and DB View: Semi-Structured and Web site as DB view of data.
      - Main Data - Web content mining has two types of view of data: IR View: Text documentation and
      - Hypertext documentation and DB View: Hypertext documentation.
      - Representation - IR View: Bag of words, n-gram term, Relational, phrases, concepts or ontology and DB View - Association rules and proprietary algorithm.
      - Method - IR View: Machine learning and statistical and DB view: Association rules and proprietary algorithm.
      - Application categories - IR View: Classification, clustering, finding extract rules and finding pattern in text and DB views Finding frequent sub structures and web site schema discovery.

2) Web Structure Mining- Web structure mining usages graph theory to examine the node and connection structure of a web site. Web structure mining can be divided into two kinds: Extract pattern from hyperlink in the web and Mining the document structure.

   Web Structure Mining Terminology:
   - View of data - Link structure
   - Main data - Link structure
   - Representation - Graph
   - Method - Proprietary algorithm
   - Application categories - Categorization and clustering

3) Web Usage Mining- It is the application of data mining techniques to determine interesting usage pattern from web data in order to recognize and superior serve the needs of web based application.

Web Structure Mining terminology:

- View of data - Link structure
- Main data - Link structure
- Representation - Graph
- Method - Proprietary algorithm
- Application categories - Categorization and clustering

E. Text Mining

The idea of text mining is to process unstructured (textual) data in order to acquire meaningful transformation. Text mining is performed on the unstructured data. See Fig. 4.

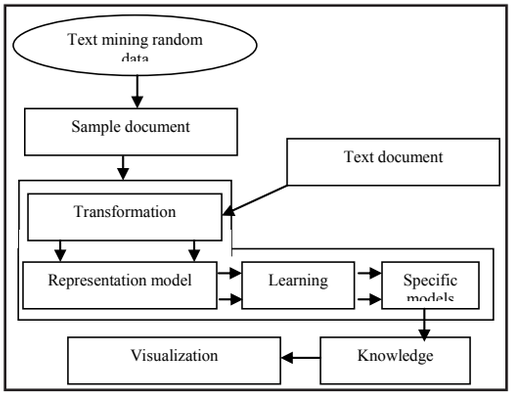


Fig. 4: Text Mining Technique

Application area of Text mining Technique:

- Security Application
- Biomedical Application
- Software Application
- Online media Application
- Business and marketing Application
- Sentiment Application
- Academic Application

II. Literature Review

Mr. Arun Kumar Sangaiah (2017) [2] found that Taxi drivers always look for strategies to locate passengers quickly and therefore increase their profit margin. In reality, the passenger seeking strategies are mostly empirical and substantially vary among taxi drivers. From the history taxi data, the top performing taxi drivers can earn 25% more than the ones with mediocre seeking strategy in the same period of time. This paper focuses on the problem of mining efficient operation strategies from a large scale history taxi traces collected over one year. Our approach presents generic insights into the dynamics of taxicab services with the objective of maximizing the profit margins for the concerned parties. We propose important metrics such as trip frequency, hot spots and taxi mileage, and provide valuable insights towards more efficient operation strategies. We analyze these metrics using techniques like Newton's polynomial interpolation and Gamma distribution to understand their dynamics. Our strategies use the real taxicab traces from the city of Changsha (P. R. China), may predict the taxi rides at different times by 90.68% per day, and increase the taxi drivers income levels up to 19.38% by controlling appropriate mileage per trip and following the route across more urban hot-spots.

Nemeth, M, Michal Conok G (2017) [3] The aim of this paper is to examine possibilities for the initial data analyses of the failure data from industrial production process. Before applying declared techniques and methods it was required to know the principle of the industrial production process itself and also to be aware of the failure data structure. Established on this, we are capable to point out interesting issues that can be further solved with KDD (knowledge discovery from databases) techniques.

Shagun Sharma, A Sai Sabitha 2016 [4] found that Data mining refers to extraction of information from huge chunks of the dataset. It's also called information mining. It is exercised in numerous fields like medicine, environment, education, crime, etc. In this research work crash investigation and analysis of the flights are done. Flight crashes may be caused due to pilot error, mechanical failure, bad weather, sabotages or human error. This research paper investigates international flight crashes since 1908 to 2009 through K-Means clustering data mining technique and cosine similarity. Clustering helps to put objects into the same group. Cosine similarity measure helps in finding similarity among different texts. The research work is done for identifying aboard/ground fatality rate with operators and location as well as to find similarity among the plane crashes. The most common reasons for plane accidents are pilot error, Mechanical failure, human error, etc. These parameters are explained as follows:

Table I: Literature Review

Name	Source of Publication	Year	Author	Technique
Efficient Mining taxi operation strategies from large scale geolocation data	Journal	2017	Mr. Arun Kumar Sangaiah.	Hotspot metric, clusteringTechniques
The initial analysis failure emerging in production process for further data mining analysis	Conference	2017	Nemeth, M, Michal-Conok G	Data mining drill down analysis

Name	Source of Publication	Year	Author	Technique
Cluster Based Best Match Scanning for large scale missing data imputation	Conference	2017	Weiquing yu, wendarg zhu, Guangyiliu	Cluster based best match scanning (CBBMS)
FlightCrash investigation technique	Conference	2016	Shagun Sharma,. A.sai sabitha	Clustering, cosine similarity, Density Based clustering
A Review Paper On Data Processing : A Critical phase in web usage mining process	Conference	2015	Dr. Sanjay Kumar Dwivedi, Bhupesh Rewet	Data Processing, Web Usage Mining, Web Structure Mining
An improved parallel K Means clustering algorithm with Map reduce	Journal	2013	Qing Liao, Fan Yang	Improved K-Means clustering algorithm

1. Pilot Error - Roughly 50% of the aircraft losses Incur due to pilot error. There are many chances for the pilots to cause errors from failing to program correctly to miscalculation of the required fuel.
2. Mechanical Failure - Despite developments in model and manufacturing standards of the aircrafts, mechanical failures account for 20% of aircraft losses.
3. Weather - Despite of having multiple electronic aids, aircrafts still struggle to function properly when the weather turns out to be unpleasant like in storms, snow and fog.
4. Sabotage - The dangers posed by sabotage are much less than many people seem to believe. Approximately 10% of aircraft losses occur due to sabotage.
5. Human Error - Mistakes can be made by humans operating when required to work for longer hours. Air traffic controllers, dispatchers, loaders, etc are some of the jobs that are operated by humans.

Qing Liao, Fan Yang 2013 [5] K-means clustering algorithm is an influential algorithm in data mining. The traditional K-means algorithm has sensitivity to the initial cluster centers, leading to the result of clustering depends on the initial centers excessively. In order to overcome this Shortcoming, this paper proposes an improved K-means text clustering algorithm by Optimizing initial cluster centers. The algorithm first calculates the density of each data object in the data set, and then judge which data object is an isolated point. After removing all of isolated points, a set of data objects with high density is obtained. Afterwards, chooses k high density data objects as the initial cluster centers, where the distance between the data objects is the largest. The experimental results show that the improved K-means algorithm can improve the stability and accuracy of text clustering.

Dr. Sanjay Kumar Dwivedi, Bhupesh Rewet (2015) [7] Path accomplishment is a critical and difficult task in the preprocessing phase of web usage mining. We mold the data preprocessing phase to achieve our goal to mine websites designed using a content management system (cms). The data preprocessing stage includes data cleaning, user empathy, session empathy, site structure and link details formation, path completion and event generation. The paper includes work on path completion by considering different types of path generated in accessing the website designed using cms and gives a novel algorithm to form the path.

III. Conclusion

The K-Mean clustering technique was used to find the clusters. And fatality for the flight crash investigation. The fatality of ground is more than aboard. The research work can be extended using other clustering techniques like Density Based, Hierarchical clustering. The summary report of the dataset is used to identify better clusters using distance measures like cosine similarity. Cosine similarity is used for finding the similarity among the crashes. Majority of the *Aeroflot* flights crashed due to other factors like error made by air traffic controller, or inexperienced crew, etc. Majority of the *Boeing* flights crashed due to hardware, technical and other issues and crashed into either mountains or ocean out of which some caught fire or suffered from mid-air collision.

The determination of the data mining to mine information from a huge data set and make over it into a operational form for additional purpose and used in business analysis. Clustering is a significant task in data analysis and data mining applications. It is the job of arrangement a set of objects so that objects in the identical group are more related to each other than to those in other groups (clusters). Clustering algorithms can be categorized into partition-based algorithms, hierarchical-based Algorithms, density-based algorithms and grid-based algorithms. Hierarchical clustering is a method of clustering which split the similar dataset by creating a hierarchy of clusters. On the basis of these one can select easily and efficiently used for future work.

References

[1] O. Oluwatuyi, and O. N. Ileri, “Air disaster and its implications in the developing countries: A case study of Nigeria,” *Modern Social Science Journal*, 2014.

[2] H. Rong, Z. Wang, H. Zhenge, C. Hu, L. Peng, Z. Ai, and A. K. Sangaiah, "Mining efficient taxi operation strategies from large scale geo-location data," *IEEE Access*, vol. 5, pp. 25623-25634, 2017.

[3] M. Nemeth, and G. Michalconok, "The initial analysis of failure emerging in production process for further data mining analysis," *2017 21st International Conference on Process Control (PC),* IEEE, 6-9 June 2017.

[4] A. S. Sabitha, and S. Sharma, "Flight crash investigation technique," *IEEE International Conference 2017*, 2017.

[5] A. Safety, *Australian Aviation Accidents Involving Fuel Exhaustion and Starvation*, 2002.

[6] Airplanes, "Statistical summary of commercial jet airplane accidents," *Worldwide Operations 2008*, 1959.

[7] S. K. Dwivedi, and B. Rawat, "A review paper on data processing: A critical phase in web usage mining process," 2015 *IEEE International Conference on Green Computing and Internet of Things (ICGCIoT),* IEEE, 8-10 Oct. 2015.

[8] Z. Nazeri, G. Donohue, and L. Sherry, "Analyzing relationships between aircraft accidents and incidents," *Proceedings of the International Conference on Research in Air Transportation*, 2008.

[9] R. L. Grossman, *et al.*, *Data Mining for Scientific and Engineering Applications*, Springer Science & Business Media, 2013.

[10] J. Srivastava, P. Desikan, and V. Kumar, "Web mining - Accomplishments and future directions," *National Science Foundation Workshop on Next Generation Data Mining*, pp. 51-56, 2002.