# Text Document Clustering Approach Using Document Similarity Measure with K-Means Algorithm

**Mr. Ashish P. Mohod[1], Mr. Najim A. Sheikh[2], Mr. Bharat S. Dhak[3], Ms. Mausami Sawarkar[4]**
[1,2,3 & 4]Asst. Professor, CSE, Priyadarshini J.L.C.E, Nagpur,RTMNU, Maharashtra, India
**Email.** – mohod.ashish@gmail.com[1], sheikhnajim4@gmail.com[2], bharat.dhak@gmail.com[3], mausami_sawarkar@rediffmail.com[4]

**Abstract:** Document clustering is way of automatic organization of documents into clusters so that documents within a cluster have high similarity in comparison to documents in other clusters. It is measuring similarity between documents and grouping similar documents together. The study of similarity measure for clustering is initially motivated by a research on automated text categorization. There was several similarity measures used for document similarity. It provides client representation and visualization of the documents; thus helps in easy navigation also. It has been used intensively because of its wide applicability in various areas such as web mining, search engines, and information retrieval. The key of organizing data in such a way is to improve data availability and to fasten data access, so that web information retrieval and content delivery on the web are improved. The main idea is to improve the accessibility and usability of text mining for various applications. By optimizing similarity measures the optimal clusters can be formed thus performance is improved.

**Key Words:** Document, optimization, system, K-means.

## 1. INTRODUCTION:

The newest answer is data mining, which is being used both to increase revenues and to reduce costs. A key enabler of data mining is the major progress in hardware price and performance. The dramatic 99 percent drop in the price of computer disk storage in just the last few years has radically changed the economics of collecting and storing massive amounts of data. At 10 Dollars per megabyte, one terabyte of data costs 10,000,000 Dollars to store. At 10 Dollars per megabyte, one terabyte of data costs only 100,000 Dollars to store! This does not even include the savings in real estate from greater storage capacities. The drop in the cost of computer processing has been equally dramatic. Each generation of chips greatly increases the power of the CPU, while allowing further drops on the cost curve. This is also reacted in the price of RAM (random access memory), where the cost of a megabyte has dropped from hundreds of dollars to around a dollar in just a few years. PCs routinely have 64 megabytes or more of RAM, and workstations may have 256 megabytes or more, while servers with gigabytes of main memory are now days is not a very rarity. While the power of the individual CPU has greatly increased, the real advances in scalability stem from parallel computer architectures. Virtually all servers today support multiple CPUs using symmetric multi-processing, and clusters can be created that allow hundreds of CPUs to work on finding patterns in the data. Advances in database management systems to take advantage of this hardware parallelism also benefit data mining. If you have a large or complex data mining problem requiring a great deal of access to an existing database, native DBMS access provides the best possible performance. The result of these trends is that many of the performance barriers to finding patterns in large amounts of data are being eliminated. Text mining is the process of discovering information in large text collections, and automatically identifying interesting patterns and relationships in textual data [3]. It is a relatively new research area, which has recently raised much interest among the research and industry communities, mainly due to the continuously increasing amount of information available on the Web and elsewhere. Text mining is a highly interdisciplinary research area, bringing together research insights from the fields of data mining, natural language processing, machine learning, and information retrieval [10]. In particular, text mining is closely related to the older area of data mining, which targets the extraction of interesting information from data records, although text mining is allegedly more difficult, as the source data consists of unstructured collections of documents rather than structured databases.

Data mining takes advantage of advances in the fields of statistics. Disciplines have been working on problems of pattern recognition and classification. Communities have made great contributions to the understanding and application of decision trees. Data mining does not replace traditional statistical techniques. Rather, it is an extension of statistical methods that is in part the result of a major change in the statistics community. The development of most statistical techniques was, until recently, based on elegant theory and analytical methods that worked quite well on the modest amounts of data being analyzed. The increased power of computers and their lower cost, coupled with the need to analyze enormous data sets with millions of rows, have allowed the development of new techniques. The key point is that data mining is the application of statistical techniques to common business problems in a fashion that makes these techniques available to the skilled knowledge worker as well as the trained statistics professional. Data mining is a tool for increasing the productivity of people trying to build predictive models.

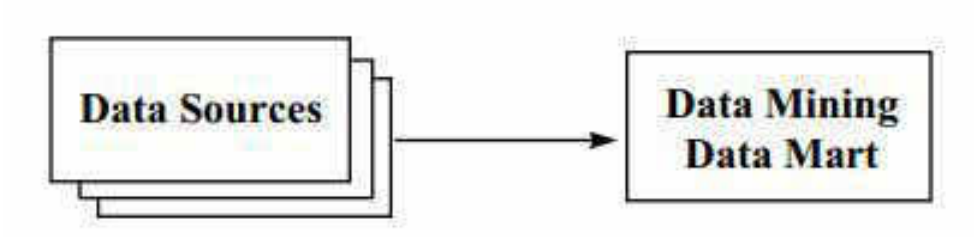


Figure 1.1.2: Data mining data mart extracted from operational databases

**1.2 Objectives:**

1. The objective of data mining is to identify valid novel, potentially useful, and understandable information in existing data. The process of data mining is used to uncover hidden or unknown information that is not apparent, but potentially useful.
2. The study of similarity measure for clustering is initially motivated by a research on automated text categorization. The objective is to improve the accessibility and usability of text mining for various applications.
3. The main objective is to propose a framework that utilizes text mining techniques in developing system. Text mining can apply intelligent methods/algorithms to extract or mine knowledge and meaningful data patterns from a large amount of unstructured texts or documents for decision-making. Therefore, it is expected that common characteristics and features from interpretations can be captured and used for standardized description to minimize the issue of subjective human judgment.
4. The aim of organizing data in such a way is to improve data availability and to fasten data access, so that web information retrieval and content delivery on the web are improved.

5. The number of clusters can be determined automatically by explicitly generating clustering for multiple values and choose the best result according to need.

**2. LITERATURE SURVEY:**

A lot of measures have been proposed for computing the similarity between two vectors. Euclidean distance is a well-known similarity metric taken from the Euclidean geometry field. Manhattan distance, similar to Euclidean distance and also known as the taxicab metric, is another similarity metric. The Canberra distance metric is used in situations where elements in a vector are always nonnegative. Euclidean distance is usually the

Default choice of similarity based methods, e.g. k-NN and k-means algorithms. Kogan et al. combined squared Euclidean distance with relative entropy in a k-means like clustering a Let d1 and d2 be two documents represented as vectors. The Euclidean distance measure is defined as the root of square differences between the respective coordinates of $d_1$and$d_2$; that is,

$$d_E U\ C(d_1; d_2) = [(d_1-d_2):(d_1-d_2)]^{1/2}$$

Cosine similarity is a measure taking the cosine of the angle between two vectors. The spherical k-means algorithm introduced by Dhillon and Modha adopted the cosine similarity measure for document clustering. Zhao and Karypis reported results of clustering experiments with 7 clustering algorithms and 12 different text data sets, and concluded that the objective function based on cosine similarity leads to the best solutions irrespective of the number of clusters for most of the data sets. Dhondt et al. adopted a cosine-based pair wise adaptive similarity for document clustering. Zhang et al. used cosine to calculate a correlation similarity between two projected documents in a low-dimensional semantic space and performed document clustering in the correlation similarity measure space.

Cosine similarity measures the cosine of the angle between $d_1$ and $d_2$ as follows: $S_{COS}(d_1,d_2) = d_1.d_2$

$$\overline{(d_1.d_1)^{1/2}(d_2.d_2)^{1/2}}$$

Pair wise-adaptive similarity dynamically selects a number of features out of $d_1$ and $d_2$ and is defined to be

$$d_{pair}(d_1, d_2) = \frac{d_{1,k}.d_{2,k}}{(d_{1,k}.d_{1,k}^{1/2})(d_{2,k}.d_{2,k}^{1/2})}$$

where $d_{i;k}$ is a subset of $d_i$ , i = 1, 2, containing the values of the features which are the union of the K largest features appearing in $d_1$ and $d_2$, respectively.

## 3. SYSTEM ARCHITECTURE:

System architecture gives brief idea about what actual working system. System architecture is the overall conceptual model that defines the structure, behavior and more views of a proposed system. An architecture description is a formal description and representation of a system, organized in a way that significantly supports reasoning about the structures of the system. System architecture comprises system components, the externally visible properties of those components, the relationships between them. It provides better plan through which efficient and effective system can be developed.

Measuring the similarity between documents is an extremely significant operation in the text as well as web data processing field. It is a symmetric measure, the difference between absence and presence of a feature is considered more essential than the difference between the values associated with a present feature. The similarity increases as the difference between the two values associated with present feature decreases [2]. The similarity decreases when the number of absence and presence features increases. The similarity measure takes the following major three cases into account:

a) The selected feature that may appear in both documents.
b) The selected feature appears in only one document.
c) The selected feature appears in none of the input documents.

### 3.1 Proposed Data similarity Measure and Document Clustering Algorithm:

Step1: Input dataset for preprocessing ie. stop word removal, steaming process.
Step2: Split each document into words and store them with their respective file path
Step3: Apply key feature selection process to find most significant terms
Step4: According to threshold value store each document key features into a newly created feature file
Step5: Create document vector for each dataset file according to present and absent feature of created feature file.
Step6: Store feature document vector into database for finding document similarity
Steps 7: Calculate similarity value using similarity function and store similarity value in database table with respective document path.
Step8: Apply K-mean algorithm to find sub-clusters based on following steps

1: Select K points as the selected initial centroids.
2: Assign all calculated similarity value points to the closest centroid.

3: Recompute the value of centroid of each cluster.

4: Repeat steps 2 and 3 until the centroids don't change to reach the desired number of clusters

Step9: Stop

## 4. CONCLUSION:

Data mining offers promising ways to uncover hidden patterns within large amounts of data. These hidden patterns can potentially be used to predict future behavior. The availability of new data mining algorithms, however, should be met with caution. First of all, these techniques are only as good as the data that has been collected. Good data is the first requirement for good data exploration. Assuming good data is available, the next step is to choose the most appropriate technique to mine the data. However, there is trade of similarity to consider when choosing the appropriate data mining technique to be used in a certain application. There are de finite differences in the types of problems that are conductive to each technique. The best model is often found by trial and error: trying different technologies and algorithms. Often times, the data analyst should compare or even combine available techniques in order to obtain the best possible results.

The process of Data mining is used to uncover hidden or unknown information that is not apparent, but potentially useful. Meaningful information lives in the form of text which is extracted from web pages. Therefore, specific pre-processing method is required in order to extract useful information from web pages. The simplest possible method for feature selection in document clustering is that of the use of document frequency to filter out irrelevant features. While the use of inverse document frequencies reduces the importance of such words, this may not alone be sufficient to reduce the noise effects of very frequent words.

## REFERENCES:


1. Yung-Shen Lin, Jung-Yi Jiang and Shie-Jue Lee,A Similarity Measure for Text Classi cation and Clustering", IEEE Transactions On Knowledge And Data En-gineering, 2013.
2. Gaddam Saidi Reddy and Dr.R.V.Krishnaiah, \Clustering Algorithm with a Novel Similarity Measure", IOSR Journal of Computer Engineering (IOSRJCE), Vol. 4, No. 6, 2012, pp. 37-42.
3. Shady Shehata, Fakhri Karray, and Mohamed S. Kamel, \An Efficient Concept-Based Mining Model for Enhancing Text Clustering", IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 10, 2010.
4. Anna Huang, Department of Computer Science, The University of Waikato, Hamilton, New Zealand, \Similarity Measures for Text Document Cluster-ing", New Zealand Computer Science Research Student Conference (NZCSRSC), Christchurch, New Zealand, April 2008.
5. H. Chim and X. Deng, \E cient phrase-based document similarity for clustering", IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 9, 2008, pp. 1217-1229.
6. Yanhong Zhai and Bing Liu, \Web Data Extraction Based on Partial Tree Align-ment", International World Wide Web Conference Committee (IW3C2), 2005
7. J. Kogan, M. Teboulle and C. K. Nicholas, \Data driven similarity measures for k-means like clustering algorithms", Information Retrieval, Vol. 8, No. 2, 2005.
8. S. Dhillon, J. Kogan and C. Nicholas, \Feature Selection and Document Clus-tering", In Berry MW Ed. A Comprehensive Survey of Text Mining,2003.
9. Syed Masum Emran and Nong Ye, \Robustness of Canberra Metric in Computer Intrusion Detection", IEEE Workshop on Information Assurance and Security United States Military Academy, West Point, NY, 2001, pp. 5-6.
10. Alexander Strehl, Joydeep Ghosh, and Raymond Mooney, \Impact of Similarity Measures on Web-page Clustering", Workshop of Arti cial Intelligence for Web Search, 2000.
11. S. Kullback and R. A. Leibler, \On information and su ciency", Annuals of Mathematical Statistics, Vol. 22, No. 1, 1951, pp. 79-86.