

# K-Mean Data Search

**Sunil Kumar.B<sup>1</sup>, Venkatesh.S<sup>2</sup>**

PG Student of CSE<sup>1</sup>, Assistant Professor of IT<sup>2</sup>  
Adithya Institute of Technology, Coimbatore

**Abstract** - In this work planned use the k-means algorithm for search engine. The principle used behind search engine is operating by crawling and indexing method. The search is performed by searching title, annotation and url of the document in the web. The k-means algorithm performs the classification operation and retrieves the data. Classification is depending on the document nature. The data segregation depends on user query. After completion of k-means operation, data transferred to the user as a three divided individual groups or frames. The groups contain web documents, portable document format (PDF) and power point presentation (ppt) of query given from the user side. The result is visualizing by the searcher.

**Index term** – k-means, Indexing, Crawling, Classification, Clustering.

## I. INTRODUCTION

Information retrieval is the process of retrieving similar data according to the user input or query. the search engine perform retrieval operation using different database. Text classification has recently become a very popular research subject in the domain of information retrieval. The purpose of text classification is to assign entries from a set of predefined categories to a file or document. A document here identified to a small piece of text. Classification may be derived from a sparse classification scheme or from a large collection of very specific content identifiers. Classification may be pointed numerically or as data and individual words. Traditionally this classification operation is calculated manually by particular domain experts. Each incoming file is observed and comprehended by the expert and then it is assign to a number of classifications selected from the set of predetermined categories. It is inevitable that a vast amount of manual effort is needed. Once the categorization scheme is learned, it can be used for classifying future documents. It involves issues commonly found in machine learning problems.

To help user to find particular or similar document from the Web, many search engines system have been created. Every search engine has a unique text database it is defined by the collection of documents that can be finding by the search engine system. In this system search engine and database will be used in interchangeable, an inverted data point for all files

in the database is opened and stored in the search engine database. For each term which can represent a meaning text or a collection of several significant words, this pointer can identify the files that have the term. Frequently, the document needed by a user is stored in different databases. As an example, consider the rack when a person wants to find research papers in particular subject domain. It is like that the particular documents are divided in a number of publisher databases. The big effort would be take for the person to find each database and identify similar documents from the retrieved files. A metasearch engine is a system that supports unique access to different existing search engines. It always does not have its own point on documents. For that it maintains the information about particular file. When a metasearch engine receives a user input, it send the input to particular local search engine and then combine the results from its particular search engines.

The documents needed by a search user are stored in different databases. example, consider the case when a search finder wants to find related documents in regarding subject domain. It is likely that the similar document papers are divided in a number of databases. Continuous effort would be needed for the search engine user to find every database and identify the similar papers from the retrieved documents related to the user needs.. A met search engine is a system that supports unique

access to different existing search engines systems. It is like that the particular documents are divided in a number of publisher databases.

The categories extraction is process of responsible for extracting the different Categories for a arrived document for newly given query. We have to make use of the different category learning method to learn the similar categories. The identified parameters in the data learning method should have been predetermined computed in the different parameter identify operation. We evaluate the classification operation performance using a new collection of files as the test collection that is varied from the training dataset. Therefore, we can analysis the classification performance by comparing the different document with the already known collection using a quality metric operation.

## II. LITERATURE SURVEY

Wai Lam, Miguel Ruiz, and Padmini Srinivasan[1], Described about Text classification has recently become a very popular research subject in the domain of information retrieval. The purpose of text classification is to assign entries from a set of predefined categories to a file or document. A document here identified to a small piece of text. Classification may be derived from a sparse classification scheme or from a large collection of very specific content identifiers. Classification may be pointed numerically or as data and individual words. Traditionally this classification operation is calculated manually by particular domain experts. Each incoming file is observed and comprehended by the expert and then it is assign to a number of classifications selected from the set of predetermined categories. It is inevitable that a vast amount of manual effort is needed. Once the categorization scheme is learned, it can be used for classifying future documents. It involves issues commonly found in machine learning problems.

Clement Yu, King-Lup Liu, Weiyi Meng, Zonghuan Wu, and Naphtali Rishe[2], described about To help user to find particular or similar document from the Web, many search engines system have been created. Every search engine has a unique text database it is defined by the collection of documents that can be finding by the search engine system. In this system search engine and database will be used in

interchangeable, an inverted data point for all files in the database is opened and stored in the search engine database. For each term which can represent a meaning text or a collection of several significant words, this pointer can identify the files that have the term. Frequently, the document needed by a user is stored in different databases. As an example, consider the rack when a person wants to find research papers in particular subject domain. It is like that the particular documents are divided in a number of publisher databases.

Ning Zhong, Yuefeng Li, and Sheng-Tang Wu[3], described about Due to the fast development of digital data complete on hand in modern years, data discovery and data mining have paying attention a vast deal of notice with an looming need for revolving such data into valuable in sequence and facts. Various applications, such as marketplace analysis and business organization, can gain by the use of their order and information extracted from a huge quantity of information. Information finding can be viewed as the method of nontrivial mining of in order from big databases, in sequence that is absolutely presented in the information, formerly unidentified and potentially practical for users. Data mining is consequently an vital step in the method of information sighting in databases.

Hassan A. Sleiman and Rafael Corchuelo[4], described about Mining data from web documents has become a study area in which fresh proposals grow out decades. This has provoked a number of researchers to work on surveys that challenge to supply an in general portrait of the lot of accessible proposals. Unluckily, no one of these surveys offer an absolute image, since they do not obtain area extractors into report. These tackle a class of preprocessors, since they help in order extractors center on the regions of a web text that have significant in a row. Through the rising difficulty of web papers, state extractors are attractive a must to take out in sequence from many websites. Further than information mining, section extractors have also start their way into information retrieval, paying interest web crawling, topic refinement, adaptive content release, and metasearch engines.

Feng Hao, John Daugman, and Piotr Zielinski[5], described about propose a fast search algorithm for

a large fuzzy database that supplies iris data with a related double arrangement. The fuzzy character of iris codes and their elevated dimensionality make a lot of fresh search algorithms, mostly relying on categorization and hashing, insufficient. The algorithm that is use in all existing community deployments of iris acknowledgment is based on a brute force exhaustive search from beginning to end a database of iris codes, look for a equal that is lock sufficient. Our fresh method, Beacon Guided Search (BGS), handle this trouble by disperse a huge number of "beacons" in the explore space. in spite of chance bit errors, iris codes beginning the similar eye are more probable tobump with the similar beacons than those from dissimilar eyes. By including the digit of collisions, BGS shrink the look forseriessignificantly with a small loss of accuracy.

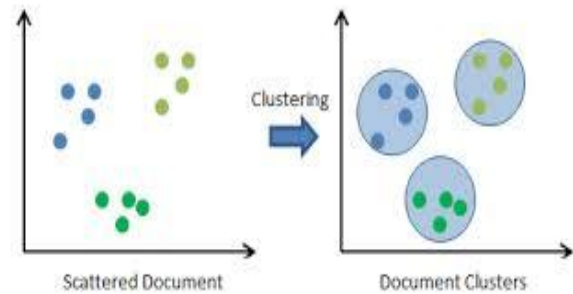
**III RELATED WORKS**

K-means clustering is a process of vector quantization originally from signal processing, that is well-liked for cluster analysis in information extracting. K-means classification aims to divide annotations into k clusters in which every examination belong to the cluster with the adjacent mean, helping as a prototype of the cluster. This outcome in a dividing of the information gap into different cells.The difficulty is computationally hard (NP-hard); but, there are well-organized heuristic algorithms that are generally engaged and join rapidly to a narrow best. These are generally related to the expectation-maximization algorithm for mixtures of Gaussian distributions through an iterative modification advance engaged by both algorithms. furthermore, they mutually use cluster centers to representation the data; though, k-means clustering tends to discover classification of equivalent spatial amount, even as the expectation-maximization method allow clusters to have unlike structure. The the majority general algorithm uses an iterative modification system. outstanding to its ubiquity it is frequently called the k-means algorithm; it is in addition referred to as Lloyd's algorithm.

$$E = \sum_{i=1}^k \sum_{p \in C_i} |p - m_i|^2$$

Where E is the sum of the square error for all elements in the data set; p is a given element; and m<sub>i</sub> is the mean of cluster C<sub>i</sub>

The below model describe clustering process of different categories into same cluster.



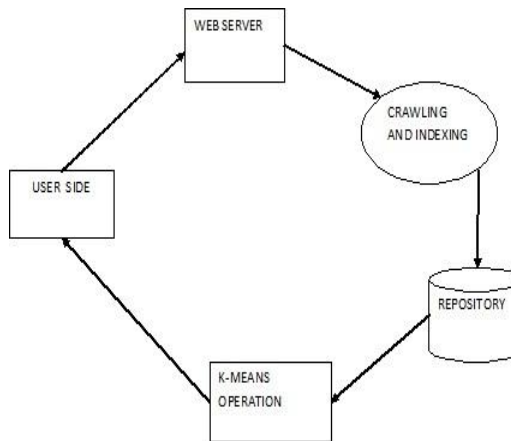
- Given k, the k-means algorithm consists of four steps:

**Algorithm:**

1. Select initial centroids at random.
2. Assign each object to the cluster with the nearest centroid.
3. Compute each centroid as the mean of the objects assigned to it.
4. Repeat previous 2 steps until no change

**IV PROPOSED SYSTEM**

In our proposed system is using the k-means method for the classification. This algorithm segregating the different classes as a individual group. The vast amount of different documents combined into groups by the nature of the document files. the retrieval of the search result is given to the user by the different class. The classes divided into web document, portable document format and power point presentation. The result of the user side is retrieved as three groups or frames. Search engine search the uniform resource locator, annotation and title of the content of individual documents.



Search engine indexing and crawling is the main process of search engine. The indexing is the process of identifying the contents relate to the user query and crawling is the method to search the millions of documents in the internet. The crawled documents or information filed in the sever system regarding priority based algorithm and send to the user. Search engine search the uniform resource locator, annotation and title of the content of individual documents. In between the collected dissimilar documents divided into individual classification method and finally information transferred to the user. The user getting result as query given to the search.

**V CONCLUSION**

This work is performed the classification operation and individual groups divided by nature of the user query. Retrieval of equivalent query search is retrieved in the browser side and user retrieving the result as a dividend frames or groups. Its minimize the search time of the user and comfortable for comparing operation between different documents. And there are a number of possibilities for extending this paper. One direction is to link this work to Web document clustering. Another direction is to apply the same model regarding performance.

**VI. REFERENCES**

[1]Wai Lam, Miguel Ruiz, and Padmini Srinivasan "Automatic Text Categorization and Its Application to Text Retrieval" 1999

[2] Clement Yu, King-Lup Liu, WeiyiMeng, ZonghuanWu and Naphtali Rishe"A Methodology to Retrieve Text Documents from Multiple Databases" 2002

[3] NingZhong, Yuefeng Li, and Sheng-Tang Wu"Effective Pattern Discoveryfor Text Mining" 2012

[4] Hassan A. Sleiman and Rafael Corchuelo"A Survey on Region Extractorsfrom Web Documents" 2013

[5] FengHao, John Daugman, and Piotr Zielin'ski "A Fast Search Algorithm for a Large Fuzzy Database" 2008

[6] C. Apte, 1:. Damerau, And S.M. Wciss, "Automated Learning Of DccisionRulcs For Text Categorization," *AcmTrairs. Liifornmtioa Systems*, Vol. 12, No. 3, Pp, 233-251, 1994.

[7] C. Buckley, G. Salton, J. Allun, And A. Singhal, "Automatic Query Expansion Using Smart: Trec-3 Llcpport," *Proc. Trec-3, IliiniText RefrievniCoif*, Pp. 69-80, 1995.

[8] W.W. Cohen And V. Singer, "Context-Sensitive Learning MethodsFor Text Catcgorizutian," *Proc. 19th Ini'lAcmSiglr Coif. RrscnrchAndDruelopnleiitIn LiforwiniionReiricunl*, Pp. 307-315, 1996

[9] C. Baumgarten, "A Probabilistic Model for Distributed Information Retrieval," *Proc. ACM Special Interest Group on Information Retrieval Conf.*, pp. 258-266, July, 1997.

[10] C. Baumgarten, "A Probabilistic Solution to the Selection andFusion Problem in DistributedInformation Retrieval," *Proc. ACMSpecial Interest Group on Information Retrieval Conf.*, pp. 246-253, Aug. 1999.

[11] N.J. Belkin, P. Kantor, E.A. Fox, and J.A. Shaw, "Combining the Evidence of Multiple Query Representations for Information Retrieval," *Information Processing & Management*, vol. 31, no. 3, pp. 431-448, May-June, 1995

[12] K. Aas and L. Eikvil, "Text Categorizations: A Survey," Technical Report Report NR 941, Norwegian Computing Center, 1999.

[13] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," *Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94)*, pp. 478-499, 1994

[14] H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections," *Proc. IEEE Int'l Forum on Research and Technology Advances in Digital Libraries (ADL '98)*, pp. 2-11, 1998.

[15] F. Hao and C. W. Chan, "Online signature verification using a new extreme points warping technique," *Pattern Recognit. Lett.*, vol. 24, no. 16, pp. 2943-2951, 2003.

[16] J. D. R. Buchanan, R. P. Cowburn, A. V. Jausovec, D. Petit, P. Seem, G. Xiong, D. Atkinson, K. Fenton, D. A. Allwood, and M. T. Bryan, "'Fingerprinting' documents and packaging," *Nature*, vol. 436, no. 28, p. 745, 2005.