Document Mining Using Relevant User Behaviour

Shailesh Choudhari [1], Dr. S. D. Choudhari [2]
MTech. CSE, Department, SBITM COE, Betul, India [1], Principal SBITM COE, Betul, India
Email - indianshailesh@gmail.com[1], choudhari.sachin1986@gmail.com[2]

Abstract: *In the digital era there is incredible growth of digital information. This information also contains supporting information with it. The supporting information is the auxiliary information which may also be useful. The auxiliary information is of the type such as citation from the scientific publications, the authorship, the co-authorship, etc. such types of side information contain tremendous amount of information. This huge amount of information may be used for performing clustering process.*

Here we deal with a principled way of performing mining process with the use of auxiliary information from the different documents. We make use of the clustering algorithm for the formation of clusters. After mining text data we search the keywords based on the user behavior so that the user may get the documents of his interest.

Key Words: *Clustering, Data Mining, Auxiliary Information*

1. INTRODUCTION:

Now a day the use of digital information is increasing tremendously. This information in the digital world is increasing to the extent that the extraction of some relevant information from this huge amount of data is becoming quite tedious. This causes an interest in creating scalable and competent mining algorithms. Till now the clustering of data in the unpolluted form is done. But to handle such large quantity of data we need to index the data according to the users need. For this we will use meta-data that is the side information that is present on most of the text documents. These meta-data correspond to various different kinds of attributes such as the origin or other information related to the origin of the document. Data such as location, possession or even temporal information may prove to be informative for mining purposes in other cases. Also once the clustering is performed a massive amount of data is received and it becomes a tedious work for user to find the documents of his/her interests from those huge clusters.

We need to check the accuracy of a system when it retrieves a number of documents on the basis of user's query input. Let the set of documents relevant to a query be denoted as "Relevant" and the set of retrieved document as "Retrieved". The set of documents that are relevant and retrieved can be denoted as "Relevant ∩ Retrieved". This is shown in the form of a Venn diagram as follows –

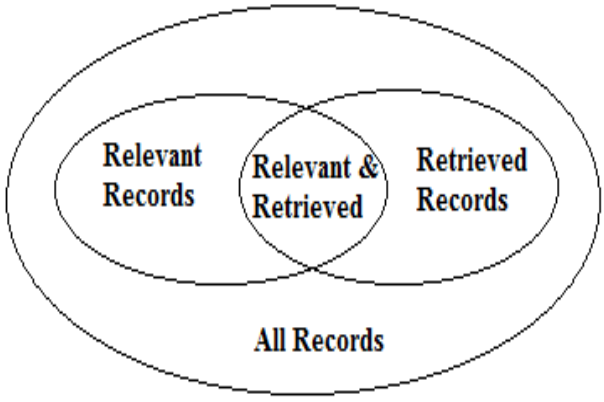

Figure 1.1: Venn Diagram

The venn diagram is shown in the above figure 1.1. The diagram shows that among all the records the relevant records and the retrieved records are shown in the diagram. For accessing the quality of the text retrieval there are three fundamental measures;

User Behavioral Search is nothing but the personalized searching [2]. The present search engines produce results that are best suited to the given query. But these engines are unaware of user's individual preferences which in turn can vary with individual interest and these interests most of the time change with individual working environment time. To provide such personalized results, user's topical preferences could be stored and utilized for the purpose.

2. RELATED WORKS

K-means clustering was explained by D. Napoleon [8]. K-means clustering also known as the centroid based clustering. This algorithm randomly selects k points as initial cluster centers. Each point from the dataset is assigned to the closest cluster and each cluster center is then recomputed as the average of points in that cluster. This is repeated until the clusters are formed.

K-means clustering attempts to maximize the intra-cluster similarity by minimizing the averaged distance between the vertices and their centroids or, equivalently, by maximizing the average similarity between the vertices and their centroids. K-means converges because the average distance between the vertices and their centroids monotonically decreases at each iterations. First, the average distance between vertices and their centroids decreases in the re-assignment step since each vertex is assigned to the closest centroid. Second, this value decreases in the recomputation step because the new centroid is the vertex for which this average distance between vertices and the centroid reaches its minimum.

One of the most popular techniques for text clustering the scatter-gather technique was introduced by D. Cutting, etal. [9]. The scatter-gather technique uses the combination of agglomerative and partitioned clustering. In the scatter-gather technique initially the system scatters the collection into small number of document groups, or clusters, and presents short summaries of them to the user. Based on these summaries, the user selects one or more of the groups for further study. The selected groups are gathered together to form a sub collection. The system then applies clustering again to scatter the new sub collection into a small number of document groups, which are again presented to the user. With each successive iteration the groups become smaller, and therefore more detailed. Scatter-gather technique is particularly helpful in situations where it is difficult or undesirable to specify query formally.

The hierarchical clustering technique was proposed by G.karypis, etal. [7]. Hierarchical clustering also known as the connectivity based clustering. Hierarchical clustering approaches attempt to create a hierarchical decomposition of the given document collection thus achieving a hierarchical structure. This clustering is based on the idea of objects being more related to the nearby objects than to the objects farther away. A cluster can be described largely by the maximum distance needed to connect parts of the clusters. Clusters can be formed based on the distances. At different distances, different clusters will be formed.

3. PROPOSED WORK

The proposed methodologies used and the proposed implementation part of the research work are described below.

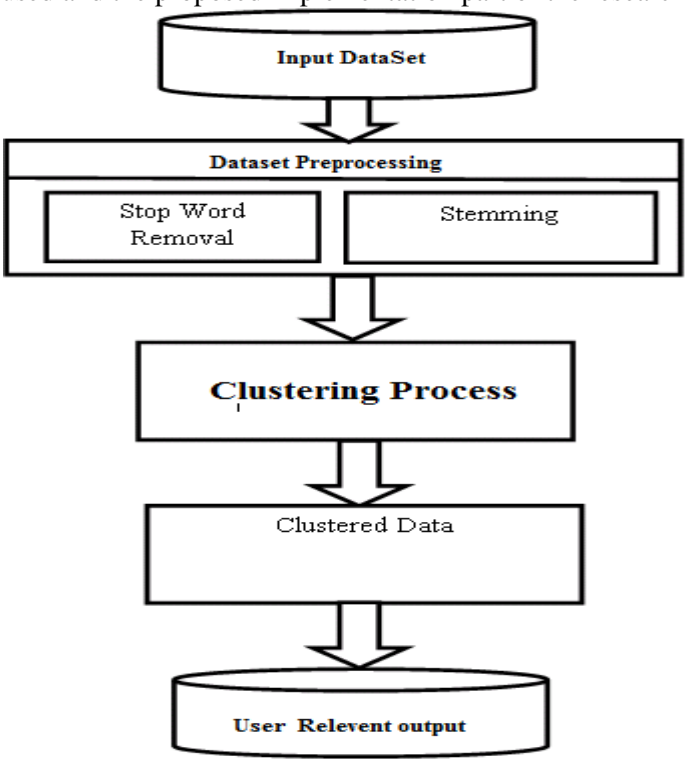

Figure 31: System Architecture

The system architecture of the project is shown in the above figure 3.1. In the proposed system first of all the objective is to collect the data on which we have to work. Once the data is collected then the pre-processing is done on the collected data set. The pre-processing includes the stop word removal technique and the stemming technique which is further explained in this section in detail. After the pre-processing, the clustering algorithm is applied. The algorithm used is explained in detail in the below section. Once the clusters are formed then the searching is done based on the user behaviour.

3.1 Steps of Proposed Algorithm

Step 1: Generally clustering is performed without using any supporting side information with the help of K means algorithm.

i. The K-means algorithm randomly selects k points as initial cluster centers.
ii. Each point from the input dataset is assigned to the closest cluster
iii. Each cluster center is then recomputed as the average of points in that cluster.
iv. Step ii and iii are repeated until the clusters are formed.

Step 2: Re-clustering is done with the help of auxiliary information it is also called as side information of the login user

3.2 Use Case Diagram

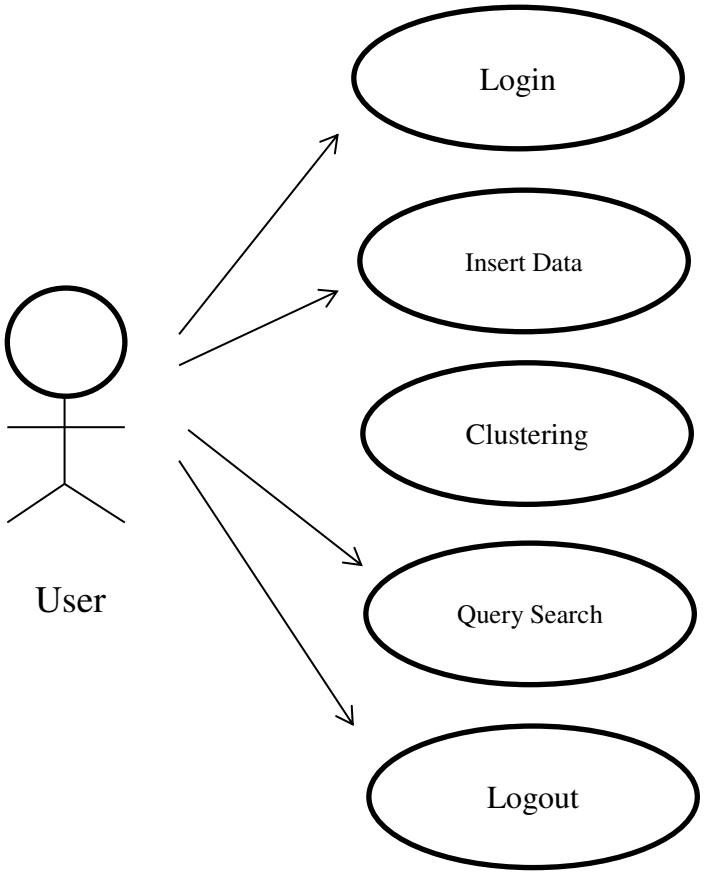

Figure 3.2: Use Case Diagram

The use case diagram for the system is shown in the figure 3.2. The User behaves as an actor for this use case diagram. Firstly the user login in the system. After the successful login of the user, the user enters the files in the system. Once the files are entered then the clustering process is performed and the clusters are formed. The user then enters the query to be searched to find the relevant documents. Then the user logs out from the system.

3.3 Sequence Diagram of the System

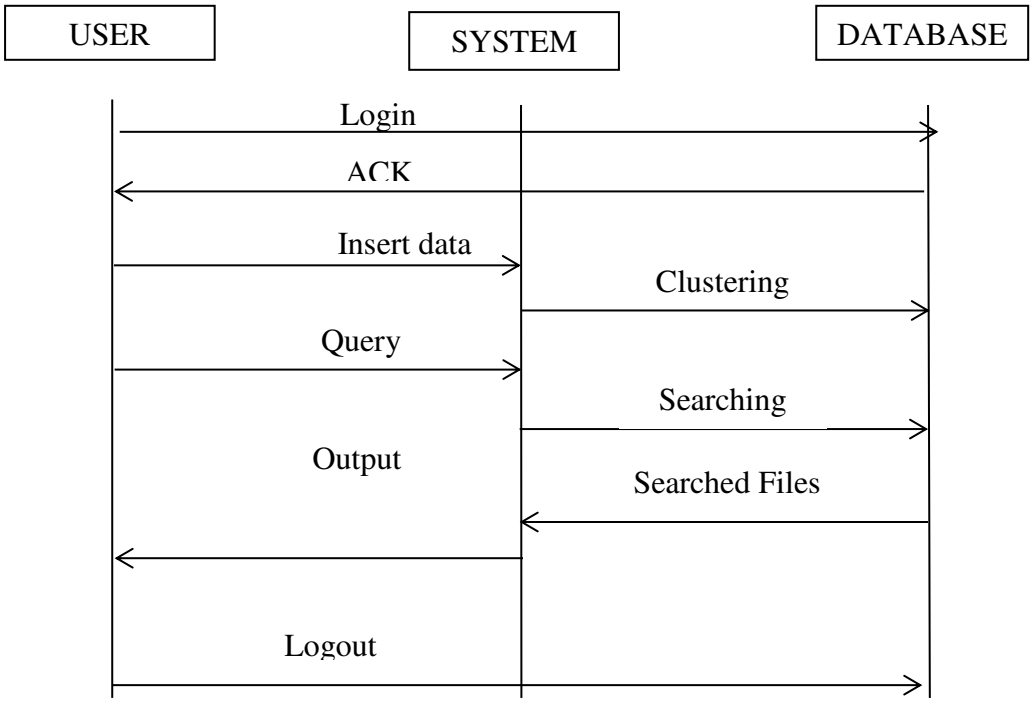

Figure 3.3: Sequence Diagram

The sequence diagram of the system is shown in the above figure 3.3 Firstly the user login by providing the user name and the password to the system. The system then verifies the user name and password with the database. If the user name and password is correct then the database gives acknowledgment to the user. After receiving the acknowledgment the user enters the documents in the system for the preprocessing purpose. Then the clustering process is done and the clustered data get stored in the database. The user then enters the query to be searched. The system searches the documents related to the query in the database. The searched files are then retrieved from the database. The output is shown to the user. Finally the user logs out from the system.

4. CONCLUSION:

The digital information from the digital era also contains supporting information which may prove to be very useful. This supporting information improves the process of clustering. In this paper we provide an efficient way of performing mining process with the use of supporting side information from the different documents. We make use of the clustering process for the formation of clusters. After mining text data we search the keywords based on the user behaviour information. This will help the user to find the desired output result which contains more relevant information.

REFERENCES:

1. Yuchen Zhao, Philip S and C. C. Aggarwal, “On the Use of Side Information for Mining Text Data”, IEEE Transactions on knowledge and data engineering, vol. 26, issue 6, June 2014.
2. Tarannum Bibi, Pratiksha Dixit Rutuja Ghule and Rohini Jadhav, “Web Search Personalization Using Machine Learning Techniques”, IEEE International Advance Computing Conference (IACC), 2014.
3. G. karypis and Ying Zhao, “Hierarchical Clustering Algorithms for Document Datasets”, Data Mining and Knowledge Discovery, vol. 10, pp. 141-168, 2005.
4. D. Napoleon and M.Praneesh, “An Efficient Numerical Methods for the Prediction of Clusters using K-means Algorithm with Bisection method for Comparing Uniform and Random Distribution Data Points”, International Journal of Innovative Research in Computer and Communication Engineering, vol. 1, issue 8, October 2013.
5. D. Cutting, J. Pedersen, J. Tukey and D. Karger, “Scatter/Gather: A cluster-based approach to browsing large document collections”, Proceedings of ACM SIGIR Conference, New York, USA, pp. 318–329, 1992.
6. Y. Gong , W. Xu and X. Liu, “Document clustering based on nonnegative matrix factorization”, Proceedings of ACM SIGIR Conference, New York, USA, pp. 267–273, 2003.

7. S. C. Gates, P. S.Yu and C. C. Aggarwal, "On using partial supervision for text categorization", IEEE Transaction Knowledge and Data Engineering, vol. 16, issue 2, pp. 245–255, February 2004.
8. P. S. Yu and C. C. Aggarwal, "A framework for clustering massive text and categorical data streams", Proceedings of SIAM Conference Data Mining, pp. 477–481, 2006.
9. J. Zhang, Q. He, K. Chang and E. P. Lim, "Bursty feature representation for clustering text streams", Proceedings of SDM Conference, pp. 491–496, 2007

WEB REFERENCES:

https://en.m.wikipedia.org/wiki/Data_mining.
https://en.m.wikipedia.org/wiki/Text_mining.
https://en.m.wikipedia.org/wiki/Information_retrieval.
http://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm.