ISSN:0975-9646



# Clustering of Biomedical Documents Using Semi supervised Clustering Method

Dr. B Bharathi[1], Anjali Vijayan[2]

[1]*Professor, Faculty of Computing, Sathyabama University, Chennai, India.*
[2]*Student, Sathyabama University, Chennai, India*

**Abstract – Clustering of biomedical document is done based on the Local content information (LC), Global content information (GC) and the Medical subject heading (MESH) – semantic information. The LC information are taken from the documents whereas the GC information from the whole MEDLINE collection. MEDLINE is a largest biomedical literature database and PubMed is an online searching service. Semi supervised spectral clustering method is used to cluster the documents. Semi supervised learning is the combination of supervised and unsupervised learning. Instance level constraints give the prior knowledge that which instances are to be grouped and which are not to be grouped. The constraints used are Must-link and Cannot-link. Must-link constraints group the documents which are similar and cannot-link group's unrelated documents.**

**Keywords – Global Content, Clustering, Supervised Learning,**

## INTRODUCTION

Data mining refers to extracting knowledge from large amounts of data. Grouping a set of physical or abstract objects into classes of similar objects is called clustering in data mining. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. In particular, document clustering means grouping of text documents into meaningful clusters in an unsupervised way. MEDLINE is a collection of biomedical documents and pubmed is an online searching service. PubMed provides an ID of the document which is user requested for. In former, only local content information were used for clustering. But latter, not only Local content information but also global content information and Mesh semantic information are used for clustering. Cluster performance is improved by the features provided by the MEDLINE documents. First, comparing the words from the title, abstract and medical subject heading pubmed provides documents related to the MEDLINE documents. Second, mesh is incorporated in the MEDLINE documents. In earlier different methods were used for clustering biomedical documents but it was ineffective for integrating different types of information. In those methods only one or two types of information is used. Then linear combination strategy has been used for clustering the biomedical documents, it was also not effective because of the limitation of representation space for combining different types of information. Thus semi supervised spectral clustering method is proposed.

Semi supervised learning is a combination of both supervised learning and unsupervised learning. A set of unlabeled objects and small amount of domain knowledge i.e., labels or pairwise constraints are the inputs of the semi supervised clustering method. The output will be partitioning of the objects into k clusters. The objectives of semi supervised clustering are maximum intra cluster similarity, minimum inter cluster similarity and provides high consistency between the partitioning and the domain knowledge. SS-COP-Kmeans, SS-HMRF-Kmeans, SS-Kernel-Kmeans and SS-Spectral-Normalized-Cuts are the semi supervised clustering algorithms used for clustering with constraints (pairwise constraints like must-link and cannot-link). K-means is a partition clustering algorithm based on iterative relocation that partitions a dataset into k clusters. Constrained k-means (SS-Kmeans) clustering is a modification to k-means clustering. Here we use constraints such as must-link and cannot-link to get the background knowledge or prior knowledge.

Spectral clustering is a graph-theoretic clustering algorithm. Semi supervised nonnegative matrix factorization is also used to incorporate prior knowledge into NMF based framework for document clustering. i.e., it includes constraints such as must-link and cannot-link. Must-link constraints work more effectively than cannot-link constraints. Here we demonstrate the performance of SSNCut using 10 data sets of MEDLINE documents with class labels and propose a method to clean the incorrect constraints.

## RELATED WORK

Khaled Hammouda and Mohamed Kamel in [1], used the core phrase key phrase extraction algorithm. The core phrase key phrase extraction method is just the opposite of the traditional keyword based clustering. This method gives more accurate representation of clusters. The algorithm first constructs a list of candidate key phrases for every cluster. It scores each candidate key phrases on the basis of its features and then ranks the candidate key phrases by score. Finally it will select the top ranking key phrases for output.

In the survey of clustering data mining techniques [2], Pavel Berkhin used the hierarchical clustering algorithm and linkage metric method. The cluster system is initialised as a set of singleton clusters in the hierarchical clustering. Then merges or splits the appropriate clusters iteratively until the stopping criterion is achieved. The similarity or dissimilarity of the cluster elements defines the appropriateness of a cluster to merge and split. From this we can understand that cluster can have similar points. To merge or split subsets of points rather than individual points, the distance between individual points has to be generalized to the distance between subsets. Such a derived proximity measure is called a linkage metric.

In paper [3], for searching the topic of the documents MALLET tool is used. It stands for Machine Learning for

Language Toolkit. It is an open source toolkit. K-means clustering algorithm is also used to cluster the documents. Partition clustering algorithm once determines all clusters. K-means clustering algorithm belongs to this category. The algorithm is as follows.

1. Choose the number of clusters, k.
2. Randomly generate k clusters and determine the cluster centers, or directly generate k random points as cluster centers.
3. Assign each point to the nearest cluster center, where "nearest" is defined with respect to one of the distance measures discussed above.
4. Recompute the new cluster centers.
5. Repeat the two previous steps until some convergence criterion is met.

In paper [4], documents that are specified with scientific texts are clustered. Clustering is done on the basis of citation contexts. A citation context is the text surrounding the reference markers used to refer to other scientific works. This citation context will provide related synonyms and vocabulary. They check how powerful this citation context is by comparing with the original document's textual representation in a document clustering.

In paper [5], the authors proposed a general framework for scalable, balanced clustering. Three steps are made to broken down the data clustering process. First, sampling the small representative subset of points. Second, clustering of sampled data. Third, populate the initial clusters with remaining data. Two steps done here are populate and refine. In populate, the points that are not sampled and hence not belong to any cluster are assigned to an existing cluster that satisfies the balancing constraints where as in refine to improve the clustering objective function iterative refinements are done.

In paper [6], here clustering is made based on similar information. The similar information put together in the same cluster where as dissimilar information will be in another cluster. This provides a way to use similarity side information to find out the clusters that reflect a user's notion of meaningful clusters. The unseen data whose similarity or dissimilarity to the training set is not known are not generalized by the instance level constraints.

In paper [7], the effects of 9 semantic similarity measures with a term re-weighting method on document clustering of pubmed document sets. The K-means clustering method points out that the term re-weighting as a method of integrating domain knowledge has some positive effects on medical document clustering.

In paper [8], authors had introduced a theoretically motivated framework for semi supervised clustering. For this, the semi supervised clustering should employ the Hidden Marcov Random Fields (HMRFs) to utilize both labeled and unlabeled data in the clustering process. This framework can be used with a number of distortion measures and it accommodates trainable measures that can be adapted to specific data sets. HMRF-K means performs this clustering in this framework. It incorporates supervision in the form of pair wise constraints in all stages of the clustering algorithm. That is initialization, cluster assignment and parameter estimation.

In paper [9], background knowledge is expressed as a set of instance level constraints. The kind of constraints used here are capable of providing prior knowledge. Prior knowledge means which instances should be grouped and which should not be grouped. The constraints used are must link and cannot link. Must link constraints provides the information that which instances should be in the same cluster whereas cannot link constraints provides the information that which all information should not be in the same cluster.

In paper [10], the aim is to find out the related documents that may be of user's interest while searching a particular document. The main task is to retrieve related MEDLINE abstracts. In other search applications, the input is a textual representation of user's information need which will be given as a short query.

In paper [11], to cluster Medline documents the semantic information of mesh thesaurus is applied by mapping documents into mesh concept vectors. To check the semantic similarity two steps are done. First, similarity between two MeSH main headings. Second, checks the similarity between two Mesh indexing sets. After the semantic similarity check, it is integrated with the content similarity and then spectral clustering is applied.

In paper [12], the general k means algorithm was enhanced by the kernel k means algorithm. A kernel function is used to enhance the k means algorithm by using an appropriate non linear from the original space to higher dimensional space. Spectral clustering uses eigenvectors of the matrix derived from the data.

In paper [13], it deals with the semi supervised learning. We will get the semi supervised learning by combining the supervised learning and the unsupervised learning. The semi supervised learning works when the knowledge on the $p(x)$ that one gain through the unlabeled data should carry the information useful in the inference of $p(y/x)$ (mathematical explanation).

## EXISTING WORK

Earlier, only one or two types of information are used. The clustering of bio medical documents was ineffective because of the difficulty faced while integrating one or two kind of information. The limitation of the earlier method is removed by enhancing the performance of MEDLINE document by linearly combining the local content and the medical subject heading (mesh) semantic information. In this method once the data are integrated then we can use any clustering methods. But this strategy also had some drawbacks. They are the true similarity what we get during the similarity check is not always a simple linear relationship between different types of similarities, the quality of the similarities in the data set should not be even for all document pairs i.e., some pairs are more important and reliable and it will be very difficult to choose a weighting configuration to balance the three or more different kinds of similarities and to integrate them.

## PROPOSED WORK

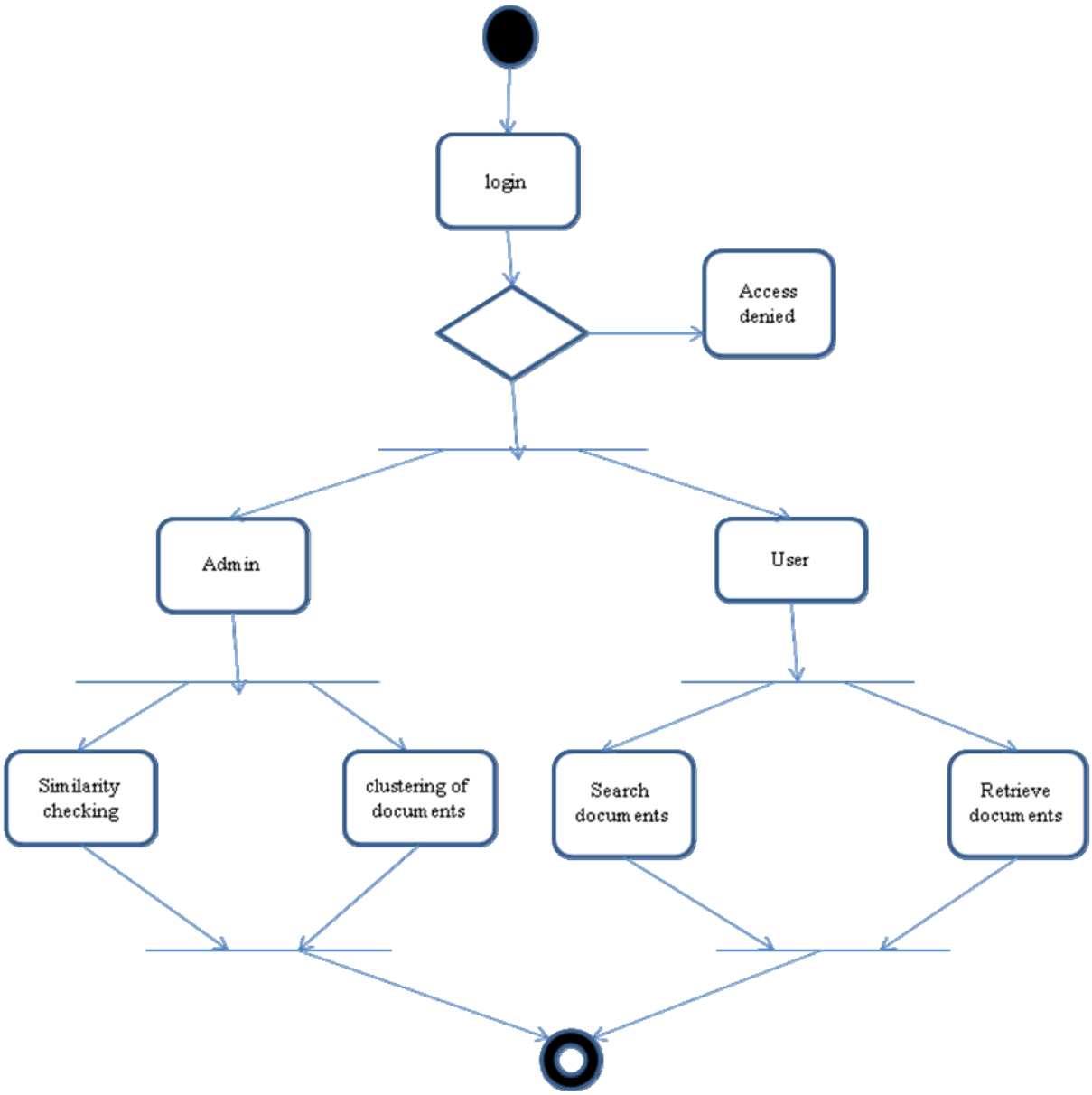
For clustering biomedical documents here we are using semi supervised spectral clustering method. There are three different types of information's we are used to cluster the

documents- Local content information (LC), Global content information (GC) and the Medical subject heading (MESH) – semantic information. The LC information are taken from the documents whereas the GC information from the whole MEDLINE collection. Must-link constraints provide the similar documents that are to be in the same cluster whereas cannot link provides the documents that are dissimilar and cannot be in the same cluster. The evaluation of each clustering method is done by comparing predicted clusters with true clusters. We understood that true clusters are not provided during the clustering process and are just used for evaluation only. There are many well-known external measures such as purity, average entropy, and mutual information for checking the performance of clustering methods. T evaluate the performance of clustering, we use the Normalized mutual information abbreviated as NMI. NMI is defined as the mutual information between the P and Q. where P and Q are the predicted class and the correct class labels. H (P) and H (Q) are the entropy of P and Q. Spectral clustering with normalized cut is clustering over the nodes in a graph. The normalized cut is the criterion that has to be reduced in the spectral clustering normalized.

$$FNC = \text{Cut}(Pk, \bar{P}k) / \text{Cut}(Pk, V)$$

Where k=1to K clusters

The system architecture for the clustering of bio medical documents shows the generation of constraints, similarity checking and the clustering of documents. Users and admin is there. User can register and login and then access the particular documents. Admin will do the similarity checking and clustering of the documents.




**System Architecture**

## CONCLUSION

Semi supervised clustering method uses three types of information to cluster the bio medical documents provided from the MEDLINE, retrieved using the online searching service pubmed. The earlier methods used to cluster the documents only based on a small amount of constraints as prior knowledge. Here we retrieve the constraints from the information such as medical subject heading (mesh) semantic and global content information. Must link constraints are more effective than the cannot link constraints. The main idea was to incorporate LC, GC and MS similarities in the document. Previously used semi supervised algorithms use only small number of constraints without noise.

## REFERENCES

[1] Khaled Hammouda and Mohamed Kamel, "Collaborative Document Clustering," 2007

[2] Pavel Berkhin, "A Survey of Clustering Data Mining Techniques," 1990

[3] Anand Karandikar, "Clustering short status messages: A topic model based approach,**"** 2010

[4] Bader Aljaber, Nicola Stoke, James Bailey and Jian Pei, "Document clustering of scientific texts using citation contexts," 2009

[5] Arindam Banerjee and Joydeep Ghosh, "Scalable Clustering Algorithms with Balancing Constraints," 2010

[6] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan and Stuart Russell, "Distance Metric Learning, with Application to Clustering with Side-Information," 2004

[7] X. Zhang, L. Jing, X. Hu, M. Ng, and X. Zhou, "A comparative study of ontology based term similarity measures on PubMed document clustering," in *Proc. DASFAA (LNCS 4443)*, 2007

[8] Sugato Basu, Mikhail Bilenko and Raymond J. Mooney, "A Probabilistic Framework for Semi Supervised Clustering," knowledge discovery and data mining, in *Proc. 10th ACM SIGKDD Int. Conf. KDD Mining*, 2004

[9] K.Wagstaff, C. Cardie, S. Rogers, and S. Schroedl, "Constrained *k*-means clustering with background knowledge," in *Proc. 18th Int. Conf. Mach.Learn*., 2001,

[10] J. Lin and W. Wilbur, "PubMed related articles: A probabilistic topic based model for content similarity," *BMC Bioinformat.*, vol. 8, no. 1, Oct. 2007.

[11] S. Zhu, J. Zeng, and H. Mamitsuka,"Enhancing MEDLINE document clustering by incorporating mesh semantic similarity," *Bioinformatics*, vol. 25, no. 15, pp. 1944–1951, Aug. 2009.

[12] Inderjit S. Dhillon, Yuqiang Guan and Brian Kulis, "Kernel kmeans, Spectral Clustering and Normalized Cuts," KDD, 2004

[13] O. Chapelle, B, Schölkopf, and A. Zien,"Semi supervised learning". Cambridge, MA: MIT Press, 2006.