# CLUSTERING OF MEDLINE DOCUMENTS USING SEMI-SUPERVISED SPECTRAL CLUSTERING

**AbinCherian[1], D.Saravanan[2], A.Jesudoss[3]**

*[1]Department of Computer Application, [2, 3]Asst. Professor, MCA, Sathyabama University, Chennai-600119*

## Abstract

*We are considering: local-content (LC) information, global-content (GC) information from PubMed and MESH (medical subject heading-MS) for the clustering of bio-medical documents. The performances of MEDLINE document clustering are enhanced from previous methods by combining both the LC and GC. We propose a semi-supervised spectral clustering method to overcome the limitations of representation space of earlier methods.*

***Keywords-*** *document clustering, semi-supervised clustering, spectral clustering*

----------***----------

## 1. INTRODUCTION

The major searching target over biomedical documents is MEDLINE, which is covering around 5600 life science journals published worldwide. We know that document clustering is grouping similar documents together and separating dissimilar documents automatically, contributes greatly to manage and organize literatures, navigate and locate searching results, and provide personalized information services. Only local-content (LC) information of documents from the data set to be clustered has been utilized for clustering.

PubMed provides a set of related articles in the whole MEDLINE collection which usually compares words from the title, the abstract, and the medical subject heading for each MEDLINE document.

## 2. EXISTING SYSTEM

There are two categories named constraint-based and distance based in the existing method. Constraint-based methods have user-provided labels or constraints to guide the algorithm towards a more appropriate data partitioning. By modifying the objective function for evaluating clustering's, it is done. Thus it includes satisfying constraints, enforcing constraints during the clustering process, or initializing and constraining the clustering based on labeled examples. An existing clustering algorithm that uses a particular clustering distortion measure is employed in the distance-based category. It is trained to satisfy the labels or constraints in the supervised data here.

### 2.1 Existing System Technique

### K-mean's clustering

1. Choose the number of different clusters, k.
2. Generate k clusters randomly and determine where the cluster centers.
3. Assign each point to the nearest cluster center, where we can define "nearest" wrt one of the distance measures discussed.
4. Recompute the new cluster centers.
5. Repeat the previous steps until some convergence criterion is met.

### 2.2 Existing System Drawbacks

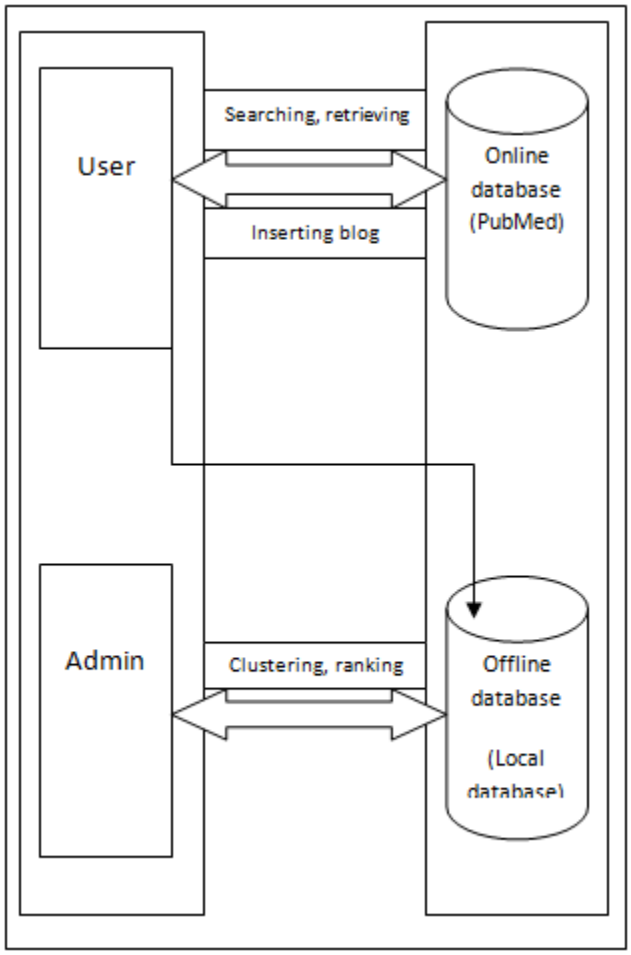
1. True similarity would not be a simple linear relationship between different similarities.
2. The quality of similarity in a data set may not be same for all document pairs. Some pairs may be more reliable and need more attention.
3. Existing system couldn't manage with a suitable weighting configuration to balance three or more different types of similarities in integrating them.

## 3. PROPOSED SYSTEM-

To improve the clustering performance, Semi supervised spectral clustering algorithms are used. The prior knowledge to improve clustering is usually provided by labeled instances or, more typically, by two types of constraints, i.e., must-link (ML) and cannot-link (CL), where ML means that the two corresponding examples should be in the same cluster and CL means that the two corresponding examples which we are considering should not be in the same cluster. We know that the Spectral clustering is a well accepted method for clustering nodes over a graph or an adjacency matrix, where clustering is

a graph cut problem that can be solved by matrix trace optimization.

### 3.1 Overall Diagram




### 3.2 Scope of the Project

By improving the performance, we have gone for alternative methods where user can search Biomedical text in our project. Usually, when user will search any text, it has to follow online databases. For searching about biomedical text, user can search documents from PubMed, Medline, PMC, Mesh, etc. These databases contain bulk amount of data. The retrieving of documents from these databases makes the performance slow. For this, we can provide option where to get documents, either from online databases or from our local database. We will make clustering of all our local database documents and can get documents from different clusters with the rank.

### 3.3 Proposed System Technique

### Semi-supervised spectral clustering

We usually use Medline, PubMed or some other databases for searching biomedical related documents. In all these databases huge number of documents are available. While retrieving those documents, performance will get slow .Hence we can retrieve some selected documents in our local database. Thus the performance could be increased. And if we go for second time search, No need to go for online Database. Get it from our local database only.

In our proposed algorithm, set of documents V (= {v1, v2, . . . ,vN}) has to be clustered. Let Sim(·, ·) be the function showing similarity between two inputs, and for example, Sim(M,M_) outputs similarity between two MeSH main headings M and M_.We denote the LC similarity matrix byWlwith the (i, j)- elementWlij, the GC similarity matrix by Wgwith the (i, j)-element Wgij, and the semantic similarity matrix by Wswith the (i, j)-element Wsij.

1. Get theurl for service given by the PubMed.
2. Right click on solution Explorer. Click add Service Reference.
3. Paste the url taken from web browser or the service url of PubMed
4. Click on go Button and in the namespace textbox, change the name as eUtils.
5. Now the proxy of service will get added in project. By using that proxy, we can call all the methods needed to retrieve the Biomedical Documents.

### 3.4 Proposed System Advantages

1. Proposed system made the most of the noisy constraints to improve the clustering performance.
2. It was viewed that ML constraints were highly powerful and CL constraints were very promising.

## 4. CONCLUSIONS

We have presented a semi supervised spectral clustering method, which can incorporate both ML and CL constraints, for integrating different information for biomedical document clustering. We have emphasized that our idea behind this project is to incorporate different type of similarities, i.e., the LC, MS and GC similarities. Semi-supervised clustering realizes this new idea, providing a more flexible framework than a method of linearly combining different similarities.

## FUTURE ENHANCEMENT

We present an application which is used to search particular biomedical documents related to our need .In this project Users are accessing biomedical documents from different clusters. As documents are well clustered and the well filtered, retrieving performance will be increased with a ranking along.

## REFERENCES

[1]. M. Krallinger, A. Valencia, and L. Hirschman, "Linking genes to literature: Text mining, information extraction,and retrieval applications for biology," Genome Biol., vol. 9, no. S2, pp. S8–S14, Sep. 2008.

[2]. D.Saravanan, Dr.S.Srinivasan, "Matrix Based Indexing Technique for Video Data ", International journal of Computer Science", 9(5): 534-542, 2013,pp 534-542.

[3]. D.Saravanan, Dr.S.Srinivasan, "Video Image Retrieval Using Data Mining Techniques "Journal of Computer Applications, Volume V, Issue No.1. Jan-Mar 2012. Pages39-42. ISSN: 0974-1925.

[4]. D.Saravanan, Dr.S.Srinivasan, " A proposed New Algorithm for Hierarchical Clustering suitable for Video Data mining.", International journal of Data Mining and Knowledge Engineering", Volume 3,

[5]. A. Rzhetsky, M. Seringhaus, and M. Gerstein, "Seeking a new biology through text mining," Cell, vol. 134, no. 1, pp. 9–13, Jul. 2008.

[6]. R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Reading, MA: Addison-Wesley, 1999. Number 9, July 2011.Pages 569

[7]. M. Lee, W. Wang, and H. Yu, "Exploring supervised and unsupervised methods to detect topics in biomedical text," BMC Bioinformat., vol. 7, no. 1, p. 140, Mar. 2006.

[8]. G. Salton and M. McGill, Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1983.

[9]. J. Lin and W. Wilbur, "PubMed related articles: A probabilistic topic based model for content similarity," BMC Bioinformat., vol. 8, no. 1, p. 423, Oct. 2007.

[10]. T. Theodosiou, N. Darzentas, L. Angelis, and C. Ouzounis, "PuReDMCL: A graph-based PubMed document clustering methodology," Bioinformatics, vol. 24, no. 17, pp. 1935–1941, Sep. 2008.

[11]. S. J. Nelson, M. Schopen, A. G. Savage, J. L. Schulman, and N. Arluk, "The MeSH translation maintenance system: Structure, interface design, and implementation," in Proc. MEDINFO, 2004, pp. 67–69.

[12]. I. Yoo, X. Hu, and I.-Y. Song, "Biomedical ontology improves biomedical literature clustering performance: A comparison study," Int. J. Bioinformat. Res. Appl., vol. 3, no. 3, pp. 414–428, Sep. 2007.

[13]. D.Saravanan, Dr.S.Srinivasan, "Data Mining Framework for Video Data", In the Proc.of International Conference on Recent Advances in Space Technology Services & Climate Change (RSTS&CC-2010), held at SathyabamaUniversity, Chennai, November 13-15, 2010.Pages 196-198.