IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Enhancing MEDLINE Document Clustering using SSNCUT With MS and GC Constraints

V. Aishwarya[*1], R. Geetha[2], R.Subash[3], G.Suresh Kumar[4]

[*1,2] Scholars, [3,4]Assistant professors, Department of Computer Science and Engineering,
Ranganathan Engineering College, Coimbatore, Tamilnadu, India
aishwaryacool30@gmail.com

Abstract

The Global content and Mesh Semantic information are considered for clustering the biomedical documents from whole MEDLER collection and Mesh Semantic information. Previously by using Semi supervised Non Negative Matrix Factorization for clustering biomedical documents are not efficient for integrating more information and inefficacious because of limited space representation for combining different analogies. To overcome this limitation a Semi supervised Normalized cut and MPCKmeans algorithm is proposed over this analogies with two constraints ML and CL constraints. And the performance of the above algorithms are demonstrated on MEDLINE document clustering.Another interesting finding was that ML constraints more effectively worked than CL constraints. We evaluate the proposed method on benchmark datasets and the results demonstrate consistent and substantial improvements over the current state. Experimental results show that integrating the semantic and content similarities outperforms the case of using only one of the two similarities, being statistically significant. We further find the best parameter setting that is consistent over all experimental conditions conducted. And finally show a typical example of resultant clusters, confirming the effectiveness of our strategy in improving MEDLINE document clustering.

Keywords: Biomedical text mining, document clustering, semi supervised clustering.

Introduction

For scientific researchers, the most important is Literature reading to vestige scientific progress and hypothesis. This MEDLINE database[1] contains over 12 million references to scientific literature with about ¾ of recent articles including an abstract of the publication. We tested the capabilities of our system to retrieve MEDLINE references which are relevant to the subject of stem cells. The Document clustering [2] is a fundamental operation used in unsupervised document organization, automatic topic extraction and the information retrieval. Clustering involves two techniques Agglomerative hierarchical clustering and k means, these are commonly used for document clustering. Initially we also believed that

agglomerative hierarchical clustering was superior to k means clustering especially for building document hierarchies and we sought to find new and better hierarchical clustering algorithm[3]. In next section ,we provide some background on the k means algorithm. And the k means clustering is method commonly used to automatically partition a data set into k groups. It proceeds by selecting k initial cluster center and then iteratively refining them respectively[4].The limitation of the existing systems is using only or two types of information and lacking effective algorithm to integrate different types of information. To overcome all above limitation the semi supervised clustering and algorithms are discussed respectively. A Semi supervised has been extensively studied in machine learning and data mining semi supervised clustering algorithm incorporate prior knowledge to improve the clustering performances. The prior knowledge is usually provided by labeled instances or more typically two types of constraints i.e., ML constraints and CL constraints where ML constraints means that the two corresponding should be in the same cluster and CL constraints means that the two examples should not be in same cluster. Here we are using a variety of clustering models. And in semi supervised clustering, a spectral clustering is a method that clusters points using eigenvectors of matrices derived from the data[5] A probabilistic topic-based model for content similarity that underlies the related article search feature in PubMed. Whether or not a document is about a particular topic is computed from term frequencies, modeled as Poisson distributions [6],[7]. Unlike previous probabilistic retrieval models, we do not attempt to estimate relevance–but rather our focus is "relatedness", the probability that a user would want to examine a particular document given known interest in another. We also describe a novel technique for estimating parameters that does not require human relevance judgments; instead, the process is based on the existence of Mesh in MEDLINE .Here the disadvantages are First there are a wide variety of algorithms that use the eigenvectors in slightly different ways. Second many of these algorithms have no proof that they will accurately compute a reasonable clustering. Previous research on cluster-based retrieval has been inconclusive as to whether it does bring improved retrieval effectiveness`s over document-based retrieval. The task of finding good clusters has been the focus of considerable research in machine learning and pattern recognition one standard approach[8] is based on generative models in which algorithm such as Em are used to learn a mixture density .A promising alternative that has recently emerged in a number of field is to use spectral methods for clustering. These suffer from several drawbacks Finally experiments of on spectral clustering analysis algorithm frequently give very poor results. Its only suitable for Particular datasets (ex : UCI). Unsupervised clustering can be significantly improved using supervision in the form of pairwise constraints, i.e., pairs of instances labeled as belonging to same or different clusters. In recent years, a number of algorithms have been proposed for enhancing clustering quality by employing such supervision. Such methods use the constraints to either modify the objective function, or to learn the distance measure. A probabilistic model for semi supervised clustering[9] based on Hidden Markov Random Fields that provides a principled framework for incorporating supervision into prototype-based clustering. The model generalizes a previous approach that combines constraints and Euclidean distance learning, and allows the use of a broad range of clustering distortion measures, including Bregman divergences and directional similarity measures We present an algorithm that performs partitioned semi-supervised clustering of data by minimizing an objective function derived from the posterior energy of the HMRF model. Experimental results on several text data sets demonstrate the disadvantages of the proposed framework. The disadvantages are in this algorithm gave one interesting problem in bioinformatics that is to improve the quality of clustering genes with unknown functions by utilizing constraints between the genes derived from domain knowledge. Not suitable for all domain applications. A semi-supervised non-negative matrix factorization framework for data clustering.[10] In Semi supervised Non Negative Matrix Factorization , users are able to provide supervision for clustering in terms of pairwise constraints on a few data objects specifying whether they “must” or “cannot” be clustered together. Through an iterative algorithm, performing a symmetric tri-factorization of the data similarity matrix to infer the clusters. Theoretically, we show the correctness and convergence of Semi supervised Negative Matrix Factorization. The correctness and convergence of the algorithm are proved by showing that the solution satisfied the KKT optimality and the algorithm is guaranteed to converge. It also prove that SS-NMF is a general and unified framework for semi-supervised clustering by establishing the relationship between SS-NMF and other existing semi-supervised clustering algorithms. Experiments performed on various publicly available datasets demonstrate the superior performance of the proposed work. Clustering based on spectral graph partitioning has emerged as a popular method over the years with applications across various domains . These methods model the data objects as vertices of a weighted graph with edge weights representing the similarity between two data objects. Clustering is then obtained by “cutting” the graph vertices into different partitions. Moreover, it shows that SS-NMF provides a general framework for semi-supervised clustering. Existing approaches can be considered as special cases of it. Through extensive experiments conducted on publicly available datasets, it demonstrates the superior performance of SS-NMF for clustering[11]. On the other hand, we compute a measure of semantic similarity between two MEDLINE documents by using their Mesh main headings and their similarities over the Mesh thesaurus, without mapping them into a common base vector, which was used in previous approaches (Yoo et al., 2006, 2007; Zhang et al., 2007).We combine the content and semantic similarities over documents, and then perform spectral clustering over the integrated similarity. Our approach contains a lot of alternatives in similarity measures and parameters to be controlled, such as the one controlling the balance between the semantic and content similarities. In our experiments, we first generate various 50 datasets of MEDLINE documents with known classes labels (biological topics). We then conduct various experiments to examine the average clustering performance over all datasets by changing alternatives and parameter values [12],[13]. Finally, we present some interesting examples of resultant clusters with different setting of calculating similarity.

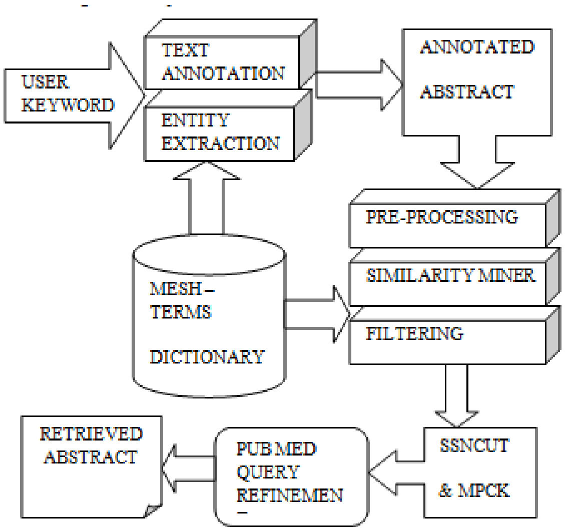

Fig 1 The whole MEDLINE Architectural overview

The Figure1 shows an architectural overview of the whole MEDLER document clustering using some of the constraints .The Text annotation is the practice and the result of adding note or gloss to a text which may include highlights, comments, footnotes, tags and links. Here similarity minor is done using linear combination method and the filtering is done in MESH indexing and finally the user required abstract is retrieved by using MESH term dictionaries and PUBMED query refinement. significant. Moreover, the performance of SSNCut and MPCK using constraints from both the MS and GC similarities is better than that using only one type of similarity, meaning that our strategy of using three types of similarities is useful in MEDLINE document clustering. Another interesting discovery is that ML constraints more effectively worked than CL constraints, partially because around 10% of generated CL constraints were incorrect, while incorrect ML constraints were only around 1%.

Proposed Work

A. Preprocessing

In this module first the files are uploaded and verify the uploaded file. The modification should be done in the pre-processing module. The modifications are all the capital letters are transformed into lower case and the stop words are removed and convert plurals into singulars. Data[14] pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values , impossible data combinations ,missing values, etc..In Pre-processing data goes through a series of following five steps

Data Cleaning: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.

Data Integration: Data with different representations are put together and conflicts within the data are resolved.

Data Transformation: Data is normalized, aggregated and generalized.

Data Reduction: This step aims to present a reduced representation of the data in a data warehouse.

Data Discretization: Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.

B.Mesh Indexing:

In this module the indexing is performed. First the documents should be stored in data set. After that we can filter the details according to the user requirements using marked up text. In mesh indexing engine it contains the list of words, stop words list, tokens, markup text. All these list form a collection and create a mesh indexed document. MESH main heading including these five major topics and they are Datastore, Filter, Sectioner**,** Lexer, Engine respectively are annoted by Each MEDLER document.

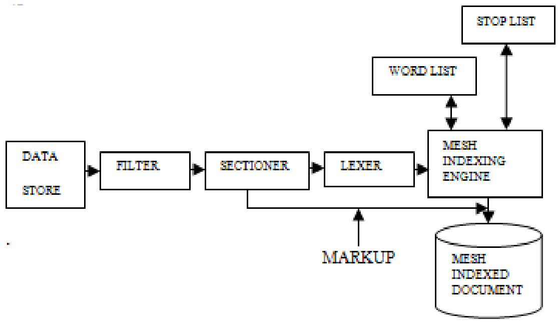

Fig 2 mesh indexing

C. Analogies:

1) Mesh Semantic Analogies:

The semantic analogies measures evaluated in this study are defined in this section and we modified some measures so that they conform to the universal definition of analogies measures are limited to the interval [0,1] and using this analogies in proceeding steps and they are Analogies between two Mesh heading and Analogies between two mesh indexing sets respectively[15],[16]. The first step describes the analogies between two nodes in a semantic network by considering four analogies and they are conceptual analogy, normalized path length, universal analogy method and combined approach. And in the second step corresponding to mesh indexing set ,each Medline documents contains a mesh main heading.

*2) **Global Content Analogies:***

This analogy can be computed between pubmed using jaccards coefficient and vector space model respectively.

D. Integrated analogies:

In this module we follow the simple approach called linear combination method for integrating mesh semantic analogy and global content analogy .The effectiveness are demonstrated by normalizing these analogy matrices.

*E. **Spectral clustering:***

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. Spectral clustering can be further divided into several variants. Normalized cut is a typical one, which minimizes the cost of intercluster edges under the constraint of the volume (the sum of node degrees) in clusters. Proposed a constrained normalized cut method, which incorporates ML constraints into the input adjacency matrix but does not consider CL constraints, which must be important for improving clustering performance. The well-accepted approach is spectral clustering which is applied to document clustering and it is similarity matrix which is equal to nodes in graph. The integrated analogies are clustered using eigen vectors. And it can be solved by matrix tree optimization and divided into normalized cut by using SS NMF algorithm. It incorporates ML and CL constraints.

*F.**Semisupervised Clustering:***

Semi supervised clustering has been extensively studied in machine learning and data mining. Semi supervised clustering algorithms incorporate prior knowledge to improve the clustering performance. The prior knowledge is usually provided by labeled instances or, more typically, 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 examples should not be in the same cluster. It generates the ML and CL constraints independently from MS and GC analogies respectively. It can clustered using normalized cut.

Experiments

In this section, we empirically evaluate the performance of our proposed method in comparison with current state of Medline document clustering. Below we will first explain our experimental setup.

*A) **Experimental Setup***

*1) **Data Sets***

The starting point of our algorithm is a set of articles associated (or believed to be associated) with a topic of interest. The system is trained with this set and therefore we define it as the *training set*. To ease evaluation of the method, we chose a subject for which the fraction of articles in the database would be neither too small nor too large of a subset of MEDLINE. In this work we used the topic *stem cells* and we took advantage of the annotation of MEDLINE [17],[18] entries with terms of the Mesh keyword hierarchy to select the training set. For this we obtained by license the complete MEDLINE database (November 2003 release, National Library of Medicine). The Mesh vocabulary contains 22,568 descriptors, and 139,000 headings called Supplementary Concept Records. An average of 10 Mesh indexing terms are applied to each MEDLINE citation by NLM indexers, who after reading the full text of the article will choose the most specific Mesh heading(s) that describe the concepts discussed. The Mesh indexing terms are organized into concept hierarchies (directed acyclic graphs) .

*2) **Runtime Analysis:***

In this paper we are using ss k means, ss nmf, semi supervised spectral clustering and porter stemming algorithm however that, practically the matrix computation are very time efficient. SS-NMF uses an iterative algorithm, and the time complexity is $O(QKN2)$, where Q is the number of iterations Similar to standard k-means, the time complexity of SS-Means is $O(QKN)$.

*3) **Evaluation Criteria:***

Two evaluation criteria are used in our experiments. First, we use normalized mutual information (NMI) to evaluate the clustering assignments against the ground-truth class labels [19]. NMI considers both the class label and clustering assignment as random variables, and measures the mutual information between the two random variables, and normalizes it to a zero-to-one range. In general, let C be the random variable representing the cluster assignments of instances, and K be the random variable representing the class labels of the instances, the NMI is computed by the following equation:

$$NMI = \frac{2I(C;K)}{H(C)+H(K)}$$

Where is the mutual information $I(X;Y) = H(X) - H(X|Y)$ between random variables X and Y .H(X) is the entropy of X,and H(X/Y) is the conditional entropy X given Y . We also note that in the early stages, the performance of the three nonrandom methods is fairly close. As we increase the number of queries, the performance advantage of our method becomes more and more pronounced. This is expected because our method make more explicit usage of the current clustering solution when selecting the queries. As we increase the number of queries, the clustering solution will become better and better, leading to more pronounced performance advantage of our method.

The Comparison graph for clustering MEDLINE documents by using SSNMF and K Means algorithms.

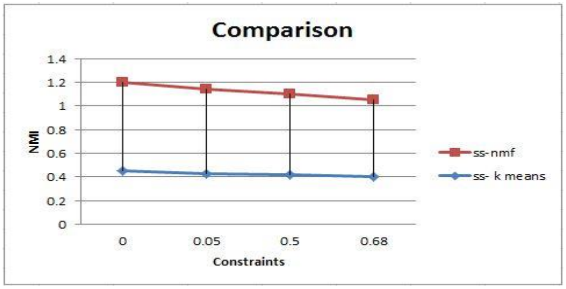

Fig 3 Comparison between SS-NMF and SS K Means

The standard spectral clustering with normalized cut with each of the three similarities or the integrated similarity by an LCM to measure the baseline performance on the data set. Through the performance of SSNMF [20] over the LC similarity with constraints created from the semantic and GC similarities, changing thresholds for obtaining ML and CL constraints. Here, we check the effect of ML constraints only first, then CL constraints only, and both types of constraints. The following graph shows the experimental result of incorrect constraints was reduced.

The idea of SL is hard constraints, which directly modify the adjacency matrix with ML and CL constraints; the weight between two corresponding instances is one for ML and zero for CL. Then, the new matrix is used for spectral clustering after normalization. Semi supervised nonnegative matrix factorization (SS-NMF) has been also developed to incorporate the ML and CL constraints similar to SL, the weight of two instances is set high for ML and low for CL.

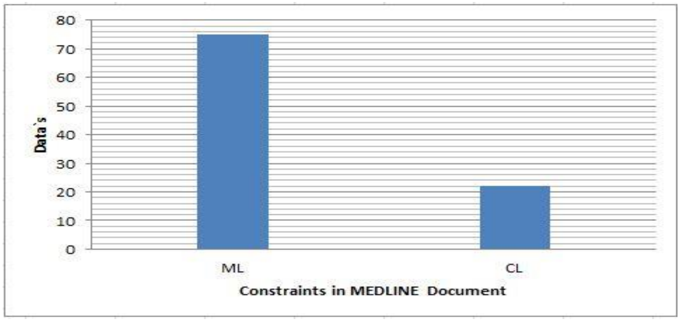

Fig 4 Performance between ML and CL

By comparing all the algorithm performance semi supervised clustering with normalized cut with ss k means, ss nmf, semi supervised spectral clustering .The below fig 5 shows that semi supervised clustering with normalized cut outperformed all other algorithms.

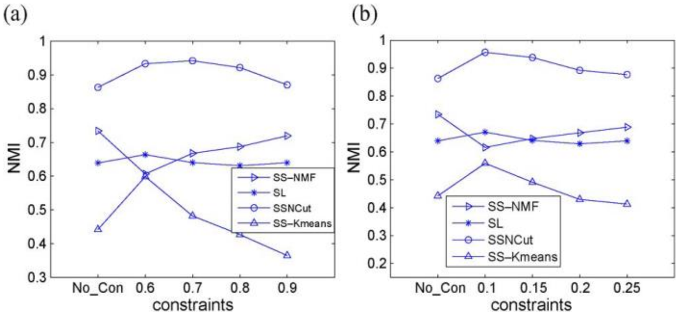

Fig 5 The NMI of algorithms with (a) ML-MS and (b) ML-GC

These result clearly indicate the advantage of semi supervised clustering with normalized cut comparing all other clustering methods respectively.

Discussion and Conclusion

The Spectral Clustering with three similarities i.e. GC constraints and MS constraints are integrated by using Linear Combination Method which is to measure only the baseline performanace.It checks the effect of ML constraints and then CL constraints, then both the ML and CL constraints. And further conduct a comparative experiment of SS-NMF and SS K Means to conform performance superiority through graph. Finally by enhancing semi supervised clustering, the MEDLINE documents are clustered by combining MS constraints and GC constraints. Here only the small number of constraints is evaluated by semi supervised clustering without noise. A new semi supervised spectral clustering method is enhanced, i.e., SSNCut, which can incorporate both ML and CL constraints, for integrating different information for document clustering. It emphasize the idea behind this paper is to incorporate three different types of document similarities, i.e., the LC, GC and MS similarities. SSNCut realizes this new idea, providing a more flexible framework than a method of linearly combining the three similarities. Once again, from these results, we can say that ML constraints were highly powerful and CL constraints were very promising, resulting in that incorporating them in SSNCut outperformed an LCM, being statistically significant. Developing a new method to clean incorrect constraints and improve the performance of SSNCut would further be an interesting future work.

References

[1] *E. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, M. Feolo,L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I.Mizrachi, J. Ostell, A. Panchenko, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J.Wilbur, E. Yaschenko, and J. Ye, "Database resources of the national center for biotechnology information," Nucleic Acids Res., vol. 38, no. 1, pp. D5–D16, Jan. 2010.*

[2] *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.*

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

[4] *R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Reading, MA: Addison-Wesley, 1999.*

[5] *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.*

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

[7] *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.*

[8] *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.*

[9] *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.*

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

[11] *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, pp. 115–126.*

[12] *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.*

[13] *D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer, "Coclustering of biological networks and gene expression data," Bioinformatics, vol. 18, no. S1, pp. 145–154, Jul. 2002.*

[14] *W. Pan, "Incorporating gene functions as priors in model-based clustering of microarray gene expression data," Bioinformatics, vol. 22, no. 7, pp. 795–801, Apr. 2006.*

[15] *D. Huang andW. Pan, "Incorporating biological knowledge into distance based clustering analysis of microarray gene expression data," Bioinformatics, vol. 22, no. 10, pp. 1259–1268, May 2006.*

[16] *M. Shiga, I. Takigawa, and H. Mamitsuka, "Annotating gene function by combining expression data with amodular gene network," Bioinformatics, vol. 23, no. 13, pp. i468–i478, Jul. 2007.*

[17] *O. Chapelle, B. Schölkopf, and A. Zien, Semi-Supervised Learning. Cambridge, MA: MIT Press, 2006.*

[18] *K.Wagstaff, C. Cardie, S. Rogers, and S. Schroedl, "Constrained k-means clustering with background knowledge," in Proc. 18th Int. Conf. Mach. Learn., 2001, pp. 577–584.*

[19] *D. Klein, D. Kamvar, and C. Manning, "From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering," in Proc. 19th Int. Conf. Mach. Learn., 2002, pp. 307–314. E. P. Xing, A. Y. Ng, M. I. Jordan, and S.*

Russell, "Distance metric learning, with application to clustering with side-information," in Proc. Adv. Neural Inf. Process. Syst., 2003, vol. 15, pp. 505–512