Medline Document Clustering with Semi-Supervised Spectral Clustering Algorithm

D.Gowthami[*1],S.Sudha[*2],V.Vishnupriya[*3],J.A.Dhinesh Joseph[#4]

[*1,*2,*3]*U.G Students, B.E CSE, Alpha College of Engg, Chennai, T.N, India.*
[#4]*Assistant Professor, Dept of CSE, Alpha College of Engg, Chennai, T.N, India.*

***Abstract*–**To clustering biomedical documents, three different types of information's are used. They are local content (LC),global content(GC) and mesh semantic(MS).In previous method only one are two types of information are cluster using Constraints and distance based algorithm. But in proposed system we used Semi Supervised clustering algorithm. It made most of the noisy constraints to improve clustering performance. The result will be highly powerful and very promising.

Keywords*–Biomedical text mining, document clustering, semi supervised clustering, spectral clustering.*

I.INTRODUCTION

Literature reading is an important approach for bio-medical researchers to trace scientific progress and generate new scientific hypothesis. The most major searching targetis MEDLINE, the largest biomedical literature database, covering around 5600 life science journals published worldwide resulting in around 21 million citations dating back to 1948,further being with PubMed (http://www.ncbi.nlm.nih.gov/pubmed/), an online searching service [1]. Mining MEDLINE for efficient knowledge discovery has become an active re-search field. In particular, document clustering, i.e., grouping similar documents together and separating dissimilar documents automatically, can greatly contribute to managing and organizing literatures, navigating and locating searching results, and providing personalized information services [4].Many studies have been carried out on MEDLINE document clustering. Traditionally, only local-content (LC) information of documents from the data set to be clustered has been utilized for clustering where each document is represented by "bag of words," resulting in a weighted vector according to the so-called vector space model [2], and then, clustering is carried out on weighted vectors. However, MEDLINE documents have some distinct features that could be utilized for enhancing the clustering performance. First, for each MEDLINE document, PubMed provides a set of related articles in the whole MEDLINE collection, which is pre computed by comparing words from the title, the abstract, and the medical subject head-ing (MeSH) using a word-weighting algorithm . Recently, the odosiou et al. have made use of this kind of global-content (GC) information for clustering MEDLINE documents .Second, most of MEDLINE documents have been annotated by the MeSH (http://www.nlm.nih.gov/mesh/). The MeSH is acont rolled vocabulary thesaurus with a set of description terms organized in a hierarchical structure where general concepts appear at the top and specific concepts appear at the bottom. Rich semantic information in MeSHs can improve the performance of clustering MEDLINE documents. modified terms in documents into MeSH concepts according to the MeSH thesaurus, showing the improvement in clustering performance under various methods, suchas k -means, bisecting k -means, and suffix tree clustering. A similar strategy was also used in term reweighting of document clustering [3]. However, this method no longer uses original texts, causing a problem that important content information in original documents may be lost. Overall, existing approacheson biomedical document clustering have two serious limitations: 1) using only one or two types of information and2) lacking effective algorithms to integrate different types of information. More recently, we have proposed an approach of linearly combining both the LC and MeSH-semantic (MS)similarities, empirically showing the performance advantage over that using only one of the two similarities .The linear combination strategy has been also used in other bioinformatics problems, such as gene clustering with multiple data (or constraints), including Gene Ontology, metabolic networks, and gene expression. In this case, once the datasets are integrated, we can use a variety of clustering models, e.g., hierarchical clustering , Gaussian mixture model ,k -medics, and Markov random fields . However, thisstrategy has roughly three underlying drawbacks in document clustering. First, the true similarity is not necessarily a simple linear relationship between different types of similarities (sources). Second, the quality of similarity in a data set may not be even for all document pairs. Some pairs are more reliable and should be paid more attention. For example, two documents with extremely high MS similarity usually reflects similar interests to be in the same cluster, while extremely low similarity might mean that the corresponding documents should not be in the same cluster. Third, it would be difficult to choosea suitable weighting configuration to balance three or moredifferent types of similarities in integrating them. Recently, semi supervised clustering has been extensivelystudied in machine learning and data mining[4].Semisu-pervised clustering algorithms incorporate prior knowledge toimprove the clustering performance. The prior knowledge isusually provided by labeled instances or, more typically, twotypes of constraints, i.e., must-link (ML) and cannot-link (CL),where ML means that the two corresponding examples shouldbe in the same cluster and CL means that the two examples should not be in the same cluster [18]–[27]. Constrainedk -means (SS-K-means) is an earlier semi supervised clustering algorithm, which was directly developed from k -means [18].In each iteration, SS-Kmeans tries to assign each instance tothe cluster with the most similar centroid, unless, at the sametime, the constraints are violated. 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. Semi supervised non negative matrix factorization (SS-NMF) has been also developed toincorporate the ML and CL constraints [26]; similar to SL,the weight of two instances is set high for ML and low for CL.Practically, a large number of constraints cannot be givenapriori, and there are no established methods to generateconstraints. Thus, in our case, by using one type of similaritiesfor examples to be clustered and the other types of similarities for constraints, semi supervised clustering can address the limitation of linear combination strategy and might improve the performance of existing methods for clustering MEDLINE documents with three different types of similarities: the LC, GC ,and MS similarities. We present a new semi supervised cluster-ing method, which we call SSN Cut, based on the normalized cut We empirically demonstrate the performance of SSN Cut by using 100 data sets of MEDLINE documents with known class labels (biological topics). Experimental results showed that SSN Cut outperformed the up-to-date linear combination strategy, as well as several well-known semi-supervised clustering algorithms, being statistically significant. Moreover, the performance of SSN Cut 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%.

II. Related Works

The problem [1] is Generalized MaxEnt aims to find a distribution that maximizes an entropy fucntion while respecting prior information represented as potential functions in miscellaneous forms of constrains and penalties. The solution is approach leads to a family of discriminative semi-supervised algorithms that are convex. The problem[2] isPrevious probabilistic retrieval models,we do not attempt to estimate relevance but rather our focus is relatedness.The solution is The pmra retrieval model was compared against bm25, a competitive probabilistic model that shares theoretical similarities. The problem[3] is Efficient access to information contained in online scientific literature collections Is essential for life science research,playing a crucial role from the initial stage of experiment planning to the final interpretation and communication of the result .The solution is The aim of the BioCreative challenge is to promote the development of such tools and to provide insight into their performance. The problem [4] is Topic detection is the task that automatically identifies topics in scientific articles based on information .The solution is Incorporating the MeSH terms and the UMLS semantic types as additional feautures enhances the performance of topic detection and the naïve bayes has the highest accuracy. The problem is Support vector machine transduction which is a combinational problem with exponential computational complexity in the number of unlabeled samples.The solution is an alternative approach introduced in which is based on a convex relaxation of the optimization problem associated to support vector machine transduction.

III. Method

A. Client

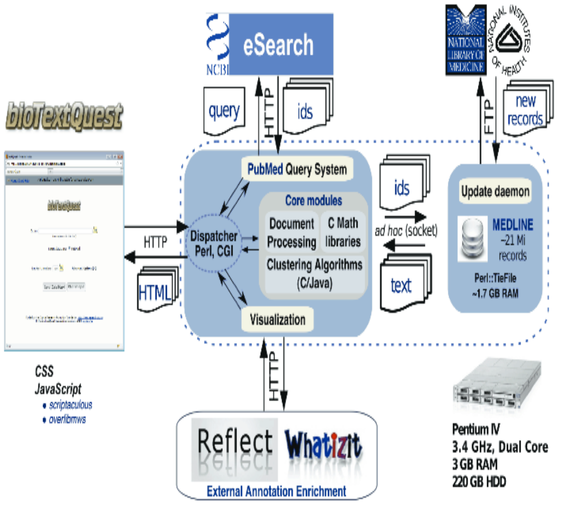
Client sends the authentication request to the admin for activating the documents from the admin or server.Clustering keywords are specified to cluster the documents. Retrieving the documents from the admin based on clustering keyword..Store all the clustered documents. $Sim(M,M□') := \sum t \in M max\ t' \in M' Sim(t,t') + \sum t' \in M' max\ t \in M Sim(t,t') ∕ |M| + |M|$

B.Similarities

1) LC Similarity: According to the well-established vectorspace model [6], each document vi is represented by a real-valued wordvector fi weighted by tfidf, which is the product of the term frequency and the inverse of their frequency in thewhole collection. We can then compute the content similarity between vi and vj by the cosine similarity of two vectors fi and fj, i.e., $W^{l}ij = (fi,fj) / |f_i| \cdot |f_j|$.

2) GC Similarity: We can compute the GC similarity between vi and vj by Jaccard's coefficient betweentheir PubMed related article sets R i and R j, i.e., W^{g}ij=|∩(R i ,Rj) | /|∩(R ,Rj) |

3) MS Similarity: We can compute the semantic similaritystep by step, according to.We first compute the similarity between two tree numbers(or two nodes in theMeSHthesaurus) t and t', i.e.,Sim (t,t') . Computing the similarity between two nodesin a semantic network is a general issue, having beenwidely discussed in the natural language processing and related fields.



C.Spectral clustering with Normalized Cut

Clustering over a similarity matrix is equal to clustering over nodes in a graph, for which a well-accepted approach is spectral clustering , which was already applied to document clustering. In spectral clustering, a popular criterion to beminimized is the normalized cut, for which the cost function of given K clusters{P1 ,...,P K } can be defined as follows: FNC =K∑k =1Cut(P$_k$,P' $_k$)/Cut(P$_k$,V)

D.SSNCuT

ML Constraint: Ji et al. incorporated ML constraintsinto the cost function of the normalized cut criterion. An MLconstraint of documents vi and vj means that viand vj musthave the same cluster preferences, i.e., that the ith and j th rows of the indicator matrix X must be the same. We then use aconstraint row vector U p=[u1p,u2p,...,unp] , where uip=1,u $_{jp}$ = − 1, and the rest of all elements are zero, and encode all ML constraints by matrix U =[U 1 ,U2 ,...,Uc]T, in which each vector is a constraints.

E.Admin

- Authentication request is granted by the admin.
- View all the documents from the stored data.
- Cluster the documents according to the keyword.

F.Data

We generate 100 benchmark data set called genomics 2005 collections, by using genomic track 2005.These 50 topics which we regarded as true clustered in our experiment, simulate real information needed in the bio medical domain and we were distributed as queries to all competing information's retrieval systems. For each topic relevant document were returned by different retrieval system, and these document were then pooled together for manual assessment by biologist. We generate benchmark dataset from these topics after removing these tiny topics to avoid small data set, after removing these tiny topics to avoid small data set because the main purpose of clustering is to find interesting cluster from large sized data and it is less meaning full to evaluate clustering methods by using smaller dataset.

Reference:

[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 liter-ature: Text mining, information extraction, and retrieval applications for biology," Genome Biol., vol. 9, no. S2, pp. S8–S14, Sep. 2008.

[3] A. 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 Bioinformatics. , vol. 8, no. 1,p. 423, Oct. 2007.

[8] T. Theodosiou, N. Darzentas, L. Angelis, and C. Ouzounis, "PuReD-MCL: A graph-based PubMed document clustering methodology,"Bioin-formatics , 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. Bioinfor-mat. 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 cluster-ing," 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, "Clustering 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 and W. Pan, "Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data," Bioinformatics, vol. 22, no. 10, pp. 1259–1268, May 2006.