College of Information Science and Technology

Utilization of Global Ranking Information in Graph-based Biomedical Literature Clustering

Xiaodan Zhang[1], Xiaohua Hu[1,2],Jiali Xia[2], Xiaohua Zhou[1],Palakorn Achananuparp[1]

[1] College of Information Science and Technology, Drexel University
3141 Chestnut street, Philadelphia, PA 19104, USA
{xzhang,thu}@ischool.drexel.edu,(xiaohua.zhou, pkorn)@drexel.edu

[2] UFSoft School of Software, Jiangxi University of Finance and Economy
Nanchang, Jiangxi, China

Abstract. In this paper, we explore how global ranking method in conjunction with local density method help identify meaningful term clusters from ontology enriched graph representation of biomedical literature corpus. One big problem with document clustering is how to discount the effects of class-unspecific general terms and strengthen the effects of class-specific core terms. We claim that running global ranking method on a well constructed term graph can identify class-specific core terms. In detail, PageRank and HITS are applied on a direct abstract-title graph to target class specific core terms. Then *k* dense terms clusters (graph) are identified from these terms. Finally, a document is assigned to the closest term graph. A series of experiments are conducted on a document corpus collected from PubMed. Experimental results show that our approach is very effective to identify class-specific core terms and thus help document clustering.

Keywords: Document Clustering, Term Graph, Global ranking

1 Introduction

It is shown that only a small portion of terms that has distinguishable power on documents clustering [9][10]. Steinbach et al. [9] argued that each document class has a "core" vocabulary of words and remaining "general" words may have similar distributions on different classes. Thus, two documents from different classes can share many general words (e.g. stop words) and will be treated similar in terms of vector cosine similarity. The ideal situation is that only distinguishable terms are used to cluster documents in a much lower dimensionality. However, to discover these distinguishable core terms is not trivial when we don't have knowledge about the document class beforehand.

HITS [6] and PageRank [8] based algorithms have been viewed as very effective approaches to calculate the global importance of a web document based on directed link information on world wide web. Moreover, LexRank [4] also showed its effectiveness on undirected graph for text summarization tasks. Therefore, in this

paper, we employ these two ranking methods on an undirected term co-occurrence graph of a given corpus to extract global important class specific core terms. However, when these algorithms are applied to a term co-occurrence graph, they face noise. The identified terms are very likely to be general terms that tend to co-occur with many "core" terms. If the noise of class-unspecific general terms is well discounted, we claim that the global ranking of class-specific core terms will be improved and only a small portion of top ranked terms will be good enough to form the initial clustering model.

We claim that this noise problem can be partially solved when the term graph is well constructed. We argue that different sections of the documents have different importance level on finding globally important class specific core terms. For example, title terms are usually more specific to the major topic of a document than that of abstract; the text of a document title usually contains much bigger percentage of topical terms than that of document abstract. Herein, a document abstract can be treated as an explanation of document title. In other words, abstract terms "cite" terms in the title. Based on this intuition, a directed abstract title term graph is constructed with abstract terms pointing to title terms. By this way, class specific core terms can get more in links from abstract terms than that of pure term co-occurrence graph.

Motivated from discussion above, a novel framework is presented to cluster a collection of documents utilizing term's global and local importance information. A collection of documents is first represented as directed title abstract term graph. PageRank and HITS based algorithms are then used to rank the terms in the graph. The top ranked terms are later clustered into k clusters. Last, a document is assigned to its closest term cluster.

Experiments are conducted on a selected PubMed document set. We make following main evaluations (1) terms' ranking on term co-occurrence graph and abstract title graph; (2) effects of different global ranking schemes; (3) quality of identified term cluster; (4) quality of identified document cluster.

Experimental results show that our approach is very effective on document clustering and can identify document clusters and core term clusters at the same time using only a small amount of distinguishable class specific core terms based on term's global and local importance information.

2 Related Work

Given the representation of documents or sentences as graph, there are some emerging works recently in text classification [5] [7] and text summarization [4].

[7] et al. represented a web document as a graph with consideration of semantic information and location of text and then extracted most frequent document Subgraphs. In the end, these document Subgraphs are used for document classification. However, in essence, this approach is equal to extract one-gram, two-gram, tri-gram, etc. from a document, and thus can not take advantage of the link information among documents and terms over the entire document set.

[5] developed an approach to cluster document by integrating term's PageRank score to documents' representation. PageRank is applied to term co-occurrence graph.

Document vectors are then represented using (PageRank Score)*IDF. The author shows it has a better performance than that of TF*IDF on clustering document by K-means. However, PageRank and IDF can be both treated as global ranking scores. Putting them together will cause information loss such as term frequency.

LexRank [4] is a PageRank based approach called power method to find globally important sentences in text summarization. It computes sentence importance based on the concept of eigenvector centrality in a graph representation of sentences. While they represent sentences as graph nodes, we take terms as graph nodes.

There are also some other works [1][2][3] that focused on how to use link information to enhance traditional content based text classification task. [2] [3] applied content based method to assign labels to part of the data and then used relaxation labeling techniques to estimate and re-estimate the class label using the hyper link information. In contrast, [1] combined content and connectivity information into a joint probabilistic model.

Although existing methods try to combine link information with content information, nonetheless, there is no approach for exploring how to identify class specific core terms using link information to facilitate initial document clustering.

3 Graph-based document clustering

3.1 Framework of the approach

The proposed approach consists of the following five main steps: (1) document representation; (2) construction of abstract-title term graph and term co-occurrence graph; (3) ranking terms according to their global importance; (4) clustering the top ranked terms to *k* clusters from term co-occurrence graph utilizing local importance information; (5) assign each document to the closest term clusters. The whole clustering process is described in the figure below.

Algorithm:

Input: an abstract title term collection graph G_AB-TI, a term co-occurrence collection graph G_CO, *k* (the desired number of clusters), p (initial # of vertices (terms) for term clustering), *M* (minimum number core terms in each cluster), Cluster quality ratio Q

Output: *k* document clusters

```
// 1: Calculating Salient Scores of vertices V
     Salience(v_i) = GlobalRanking_G_AB-TI(v_i)(Eq (1), Eq(2))
//sort V in the descending order of Salience(v)using abstract-title
graph G_AB-TI
Sort(V,Salience(v),des)

// 2: Detecting k core term clusters from Graph G_CO
Do{
    For(i=1; i<=numOfNodesForClustering; i++){
       Get free cluster Ck from k free cluster pool
       Ck.add(Ti) //add term Ti to cluster Ck
       Check in_Cluster_Degree for all terms Ti_list that have edge
 with Ti //refer to Eq(3)
       Sort Ti_list descending
       For(j=size(Ti_list);j>=1,j--){
```

```
            //check cluster quality
           if (In_cluster_degree(Tj)/In_cluster_degree(Ti)>=Q){
               cutoff_point = j;
                break loop;
                }
          }
     }
     Add terms over cutoff point to cluster Ck
     If(numOfTerm(Ck)<M){
        remove all terms from cluster Ck
        put back Cluster Ck to free cluster Pool
       }
}While reach the number of K Cluster
// 3: Assign document to closest term cluster
Assign remaining top ranked terms to K cluster by Max (Eq.(4))
Match Document To Term Cluster by Max (Eq.(5))
```

Fig. 1. Clustering algorithm

3.2 Document representation

In biomedical domain, it is very common that a concept has more than one synonym and is composed of more than one word, which makes it very necessary to represent document using appropriate biomedical ontology.

MeSH Ontology. Medical Subject Headings (MeSH) [www.nlm.nih.gov./mesh] mainly consists of the controlled vocabulary and a MeSH Tree. The controlled vocabulary contains several different types of terms, such as Descriptor, Qualifiers, Publication Types, Geographics, and Entry terms. Among them, Descriptors and Entry terms are used in this research because they are the only terms that can be extracted from documents. Descriptor terms are main concepts or main headings. Entry terms are the synonyms or the related terms to descriptors. For example, “Neoplasms” as a descriptor has the following entry terms {“Cancer”, “Cancers”, “Neoplasm”, “Tumors”, “Tumor”, “Benign Neoplasm”, “Neoplasm, Benign”}. MeSH descriptors are organized in a MeSH Tree, which can be seen as the MeSH Concept Hierarchy. In the MeSH Tree there are 15 categories (e.g. category A for anatomic terms), and each category is further divided into subcategories. For each subcategory, corresponding descriptors are hierarchically arranged from most general to most specific.

Mesh Descriptor Term extraction. While processing an abstract, we map terms in each document to the Entry terms in MeSH and then maps the selected Entry terms into MeSH Descriptors to handle the synonyms. In this way, synonyms of a given MeSH descriptor are assigned a unique ID.

We create one stop term list for MeSH based on the analysis of PubMed documents from 1994-2004 using Zipf law [12]. Based on the stop term list, we exclude some MeSH terms that are too general (e.g. HUMAN, WOMEN or MEN) or too common in MEDLINE articles (e.g. ENGLISH ABSTRACT or DOUBLE-BLIND METHOD).

3.3 Global Ranking and Term Graph construction

PageRank [8] is one of methods Google uses to determine a page's relevance or importance. The beauty of the method is that it integrates social reference knowledge into the page ranking procedure. Given a web page in a network, the PageRank score of this page is defined as follows:

$$PR(p_i) = \frac{1-d}{N} + d \cdot \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \quad (1)$$

where N is the number of pages under consideration, $M(p_i)$ is a set of pages that link to p_i, $L(p_j)$ is the number of outbound links on page p_j, and d is damping factor.

Hypertext Induced Topic Selection (HITS) [6] is a link analysis algorithm that rates Web pages by their authority and hub values. Authority value estimates the value of the content of the page; hub value estimates the value of its links to other pages. These values can be used to rank Web search results.

Let N be the set of nodes in the neighborhood graph. For every node n in N, let $H[n]$ be its hub score and $A[n]$ its authority score. Initialize $H[n]$ and $A[n]$ to 1 for all n in N. While the vectors H and A have not converged:

$$\text{For all n in N,} \quad A[n] := \sum_{(n',n)\in N} H[n'] \quad (2)$$

$$\text{For all n in N,} \quad H[n] := \sum_{(n,n')\in N} A[n']$$

Normalize the H and A vectors.

Table 1. The PageRank score of title and abstract terms of a document

PageRank score of Title term		PageRank score of Abstract term	
Gout	5.53E-06	Gout	5.53E-06
		Association	3.93E-06
		Incidence	3.47E-06
		Epidemiology	2.52E-06
		Pain	1.68E-06
		Hyperuricemia	1.20E-06
		Arthritis, Gouty	1.18E-06
		Arthritis	4.10E-07
		Algorithms	2.14E-07
		Obesity	1.62E-07
		Patient Education	1.52E-07
Arthritis, Gouty	1.18E-06	Diabetes Mellitus	1.37E-07
		Hyperlipidemia	8.97E-08

The success of PageRank and HITS lies on calculating stationary distribution on a directed social link graph. However, for a term co-occurrence graph, it does not have this information, i.e., there is no reference relationship between terms. We can build

symmetrical term co-occurrence graph where term co-occur with each other when they appear in the same document. However, ranking terms on pure term-concurrence graph can be very likely to assign general terms high ranking scores because they tend to co-occur with many other terms. Thus, how to let down the ranking of general terms will be crucial to let up the ranking of class-specific core terms. We claim that the different section of a document can provide the directed reference information like hyperlink environment. For example, let abstract terms link to title terms, so the link can be taken as a scenario in which abstract terms give a detailed explanation of title terms by referencing title terms. Since title terms usually contain much bigger percentage of topical terms than that of abstract, these topical terms can receive much higher score than within term co-occurrence graph as shown in table 1. In this way, a corpus level abstract-title term graph is built with abstract terms pointing to title terms.

3.4 Term clustering by local density

We assume that the top ranked terms will form several dense areas within the co-occurrence term graph.

$$In\_Cluster\_Degree(t_i, C_k) = Num(Edge_{t_i, C_{kj}}) \tag{3}$$

As shown in Figure 1, the term clustering process starts from the top ranked globally important term. For example, if the starting term is “Hepatitis B, Chronic”, the algorithm will take all terms that connect to “Hepatitis B, Chronic” as a candidate cluster including itself. Then each cluster member’s in cluster degree will be calculated and then sorted in descending order. The algorithm will start from the term with the least in cluster degree and calculating whether its ratio with the highest one is over threshold Q. If it is so, the algorithm will keep the terms over threshold as a core term cluster. If this core term cluster has enough number of terms, a new core term cluster is formed. Otherwise, the algorithm will skip this term and grow core term cluster from the second top ranked term. If the second is already included in the cluster, then it will start from the third and so on. As discussed earlier, each class only contains a small number of class specific core terms. We keep only a few higher ranked terms in the core cluster and remove lower ranked terms to the pool of reassignment. To guarantee the high quality of initial term cluster, we set the minimum number of core terms in each cluster to 3 and the quality ratio Q as 0.8.

The remaining top ranked terms are assigned to K term clusters according to its In_Cluster_Frequency(ICF) that is the number of edges connecting it to cluster members with edge weight counted (please refer to Fig.1for details):

$$In\_Cluster\_Freq(t_i, C_k) = \frac{\sum Weight(Edge_{t_i, C_k j})}{Num(Edge_{t_i, C_{kj}})} \tag{4}$$

where $Weight(Edge_{t_i, C_k j})$ is the edge weight between term *i* and term *j* in cluster C_k and $Num(Edge_{t_i, C_{kj}})$ is the number of edges. In this study, we use term co-occurrence as edge weight. This is also extensible to other types of weight.

3.5 Document clustering

After core term clusters are identified, each document is assigned to its closest cluster:

$$DocClusterCloseness(d_i, C_k) = \sum In\_Cluster\_Freq(d_{i,t_j}, C_k) \quad (5)$$

where d_{i,t_j} is the term t_j in document d.

4 Experimental results

4.1 Document sets

For the extensive experiments, we collect document sets about various diseases from PubMed, which is the web interface to MEDLINE. We use "MajorTopic" tag along with the disease MeSH terms as queries to PubMed. In this way, 10 document classes are collected for the experiments (see table 2).

Table 2. The document sets and their sizes

Document Sets	#.of Docs
Gout	642
Chickenpox	1,083
Raynaud Disease	1,153
Jaundice	1,486
Hepatitis B	1,815
Hay Fever	2,632
Kidney Calculi	3,071
Age-related Macular Degeneration	3,277
Migraine	4,174
Otitis	5,233

4.2 Term's Global Ranking

Table 3. Top ten terms ranked by PageRank

Abstract-Title Term Graph (ATTG)	Term Co-occurrence Graph (TCG)
Otitis	Patients
Migraine Disorders	Therapeutics
Patients	Disease
Therapeutics	Child
Child	**Otitis**
Macular Degeneration	**Migraine Disorders**
Infection	Time
Chickenpox	Infection
Hepatitis B, Chronic	Serum
Kidney Calculi	Role

Since class-unspecific general terms co-occur frequently with many other terms, how to reduce the effects of common terms will contribute to discover distinguishable

class-specific core terms. We evaluate the impacts of term graph construction on term's global ranking. As shown in the table 3, the ATTG schemes has six class specific core terms, while the TCG scheme contains only two core terms. Obviously, PageRank algorithm assigns higher weight to class specific core terms on ATTG than on TCG, which indicates that this representation is very effective on discounting class-unspecific general terms.

4.3 Term Clustering Evaluation

We evaluate the quality of term cluster by whether it contains class name related core terms (Table 2). As shown in table 4, nine out of ten semantic related and graphical connected term graphs (clusters) (containing class name related terms) are identified through our core term cluster identification algorithm. This indicates our method can create initial cluster models with high quality.

Table 4. Term cluster identified by our algorithm using PageRank

	Term cluster(corresponding class name)
1	Kidney Calculi, Shock, Lithotripsy (**Kidney Calculi)**
2	Macular Degeneration, Visual Acuity, Vision (**Macular Degeneration**)
3	Chickenpox, Viruses, Herpesvirus 3, Human**(Chickenpox)**
4	Migraine Disorders, Epilepsy, Women **(Migraine Disorders)**
5	Otitis Media with Effusion, Otitis, Observation **(Otitis)**
6	Hepatitis B, Chronic, Lamivudine, Hepatitis B virus, Hepatitis B, Antigens(**Hepatitis B**)
7	Kidney Calculi, Calcium Oxalate, Organization and Administration (**Kidney Calculi**)
8	Jaundice, "Jaundice, Neonatal", Bilirubin, Life (**Jaundice**)
9	Rhinitis, Pollen, Immunotherapy (**Hay Fever**)
10	Macular Edema, Cystoid, Visual Acuity, Edema (**Macular Degeneration**)

4.4 Document Clustering Evaluation

Cluster quality is evaluated by three extrinsic measures, *purity*[13], *entropy*[13], *F-measure*[13] and *normalized mutual information (NMI)* [11].

To test two schemes of term graph construction on document clustering, we run global ranking including PageRank and HITS on both term co-occurrence graph and abstract title term graph. The effects of the number of nodes used for document clustering are also compared. From table 5,6,7,8, we can see that (1) PageRank and HITS have very similar performances; and PageRank is slightly better; (2) abstract-title scheme is better than co-occurrence scheme when the terms used for clustering are very few. This is expected because abstract-title scheme give class-specific core terms higher ranking than term co-occurrence graph (Table 3); (3) when all the terms are used for clustering, there is not much difference between ATTG and TCG scheme because Lower ranked terms have the same chance to be grew into a cluster as top

ranked terms; (4) the most promising result is with ATTG scheme; only 100 terms are enough for clustering the entire document set into high quality document clusters. It is worth noting that 200 instead of 100 are assigned for TCG scheme, because 100 terms are not enough for our algorithm to identify core term clusters. We also compare clustering results to that of spherical K-mean clustering using TF*IDF as document representation. Our clustering performance is slightly worse than K-mean (Entropy: 0.40, F-measure:0.754, Purity: 0.889 and NMI: 0.755). However, our algorithm (including term ranking, core term cluster identification and document matching) performs more efficiently. On a Laptop PC with Duo core 1.83GHZ, 1GB memory, and 80GB hard drive setting, our algorithm usually finishes within 20 seconds, while spherical K-means costs more than 30 seconds. Spherical K-means needs to iteratively re-estimate distance between each document and cluster center. However, as long as the core term clusters are identified, our algorithm can determine documents' class labels by one time calculation. Moreover, our approach can identify distinguishable class-specific term clusters which can serve as the interpretation of clustering results, thus, our contribution is beyond the document clustering itself.

Table 5. PageRank on Abstract-Title term graph

Terms for clustering	Entropy	F-measure	Purity	NMI
100	0.500	0.811	0.885	0.661
200	0.620	0.772	0.850	0.640
All	0.633	0.763	0.840	0.654

Table 6. PageRank on term co-occurrence graph

Terms for clustering	Entropy	F-measure	Purity	NMI
200	0.780	0.737	0.812	0.581
300	0.731	0.758	0.825	0.610
All	0.646	0.764	0.839	0.650

Table 7. HITS on Abstract-Title term graph

Terms for clustering	Entropy	F-measure	Purity	NMI
100	0.522	0.799	0.785	0.641
200	0.656	0.742	0.760	0.623
All	0.641	0.763	0.775	0.642

Table 8. HITS on term co-occurrence graph

Terms for clustering	Entropy	F-measure	Purity	NMI
200	0.801	0.699	0.724	0.563
300	0.742	0.706	0.782	0.602
All	0.677	0.752	0.801	0.631

5 Conclusions and Future Work

In this paper, we present a framework of graph-based document clustering which utilizes both global and local term importance information on not only clustering documents but also identifying class-specific core term clusters. We mainly discussed two schemes of graph construction: term co-occurrence graph and abstract-title graph

and how documents are clustered based on term global and local importance from these two schemes. The main findings are: (1) the identified core dense term clusters (graphs) can be both helpful for document clustering and clustering results interpretation; (2) abstract-title graph scheme can help identify more class-specific core term clusters than term-occurrence graph scheme, which indicates encoding document section importance information to term graph can help give class-specific core terms higher ranking; (3) while performing more efficiently, our algorithm is comparable to top clustering algorithm such as spherical K-means. For our future work, we would like to evaluate our approach on other applications such as text classification and summarization; also we will extend our research to general domain.

Acknowledgments. This work is supported in part by NSF Career Grant IIS 0448023, NSF CCF 0514679, PA Dept of Health Tobacco Settlement Formula Grant (No. 240205 and No. 240196), and PA Dept of Health Grant (No. 239667).

References

[1] Angelova, R. and Weikum, G. Graph-based text classification: learn from your neighbors. SIGIR 2006: 485-492

[2] Charkrabarti, S., Dom, B.E., and Indyk, P. Enhanced hypertext categorization using hyperlinks. In SIGMOD'98, 307-318.

[3] Cohen, W.W. and Hofmann, T. The missing link—a probabilistic model of document conent and hypertext connectivity. In NIPS 13, 2001.

[4] Erkan, G., Radev, D.R.: LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. J. Artif. Intell. Res. (JAIR) 22: 457-479 (2004)

[5] Hassan, S. and Banea, C. Random-Walk TermWeighting for Improved Text Classification, Workshop on TextGraphs, at HLT-NAACL 2006, pages 53–60,

[6] Kleinberg, J. Authoritative sources in a hyperlinked environment. In Proc. Ninth Ann. ACM-SIAM Symp. Discrete Algorithms, pages 668-677, ACM Press, New York, 1998

[7] Marklv,A Last. M and Kandel A, "Model-based classification of web documents represented by Graphs", proceedings of WebKDD 2006 workshop on knowledge discovery.

[8] Page, L. and Brin,S., Motwani, R., and Winograd,T. The PageRank citation ranking:Bringing order to theWeb. *Technical report*, Stanford Digital Library Technologies Project, 1998.

[9] Steinbach, M., Karypis, G., and Kumar, V. *A Comparison of Document Clustering Techniques*. Technical Report #00-034. Department of Computer Science and Engineering, University of Minnesota, 2000.

[10] Wang, B.B., McKay, R I., Abbass, H.A., Barlow M. Learning Text Classifier using the Domain Concept Hierarchy. In *Proceedings of International Conference on Communications, Circuits and Systems 2002*, China.

[11] Zhong, S., and Ghosh, J. A comparative study of generative models for document clustering. *Proceedings of the workshop on Clustering High Dimensional Data and Its Applications in SIAM Data Mining Conference*, 2003.

[12] Zipf, G.K. *Human Behaviour and the Principle of Least-Effort*, Addison-Wesley, Cambridge MA, 1949

[13] Zhao, Y. and Karypis, G. *Criterion functions for document clustering: experiments and analysis*, Technical Report, Department of Computer Science, University of Minnesota, 2001.