

Integrating Biomedical Literature Clustering and Summarization Approaches using Biomedical Ontology

Illhoi Yoo

Department of Health Management
and Informatics, School of Medicine,
University of Missouri-Columbia,
Columbia, MO 65211, USA
yooil@health.missouri.edu

Xiaohua Hu

College of Information Science and
Technology, Drexel University,
Philadelphia, PA, 19104, USA
thu@cis.drexel.edu

Il-Yeol Song

College of Information Science and
Technology, Drexel University,
Philadelphia, PA, 19104, USA
song@drexel.edu

ABSTRACT

We introduce a method that integrates biomedical literature clustering and summarization using biomedical ontology. The core of the approach is to identify document cluster models as semantic chunks capturing the core semantic relationships in the ontology-enriched scale-free graphical representation of documents. These document cluster models are used for both document clustering on document assignment and text summarization on the construction of Text Semantic Interaction Network (TSIN). Our experimental results show our approach is superior to traditional approaches including Bisecting K-means as a leading document clustering approach in terms of cluster quality and clustering reliability. In addition, our approach provides concise but rich text summary in key concepts and sentences.

Categories and Subject Descriptors

H.3.3 [INFORMATION STORAGE AND RETRIEVAL]:
Information Search and Retrieval – *Clustering*.

General Terms

Algorithms, Experimentation, Theory

Keywords

Ontology, Document Clustering, Text Summarization, Scale-Free Network, MEDLINE

1. INTRODUCTION

A huge amount of textual information has been produced and collected in text databases or digital libraries for decades because the most natural form to store information is text. For example, MEDLINE, the largest biomedical bibliographic text database, has more than 16 million articles and more than 10,000 articles are weekly added to MEDLINE.

In order to tackle this pressing text information overload problem, document clustering and text summarization together have been

used as a solution. This is because document clustering enables us to group similar text information and then text summarization provides condensed text information for the similar text by extracting the most important text content from a similar document set or a document cluster. For this reason, document clustering and text summarization can be used for important components of information retrieval system.

In this paper, we introduce a method that integrates biomedical literature clustering and summarization using biomedical ontology MeSH. We claim that integrating document clustering and text summarization is required because a set of documents are usually multiple-topics. For this reason text summarization does not yield high-quality summary without document clustering. On the other hand, document clustering is not very useful for users to understand a set of documents if the explanation for document categorization or the summaries for each document cluster is not provided. In other words, document clustering and text summarization are complementary. This is the primary motivation for integrating document clustering and text summarization.

The rest of the paper is organized as follows. Section 2 surveys the related works. In Section 3, we propose a novel graph-based document clustering approach that uses domain knowledge in an ontology and text summarization using Text Semantic Interaction Network using the semantic relationships in the document cluster model. An extensive experimental evaluation on MEDLINE articles is conducted and the results are reported in Section 4. Section 5 concludes our paper.

2. THE PROPOSED APPROACH

We present a novel coherent document clustering and summarization approach, called *Clustering and SUMmarization with GrAphical Representation for documents (CSUGAR)*. For the detailed discussion on MeSH ontology, refer to [8].

2.1 Clustering and Summarization with Graphical Representation (CSUGAR)

The proposed approach consists of two components, document clustering and text summarization. Each step is discussed in detail below. Note the steps 1 to 3 correspond to document clustering and the steps 4 to 6 correspond to text summarization.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

TMBIO '06, November 10, 2006, Arlington, Virginia, USA.

Copyright 2006 ACM 1-59593-526-6/06/0011...\$5.00.

Step1 - Ontology-enriched Graphical Representation for Documents through Concept Mapping

The idea of the use of ontology-enriched graphical representation for documents for document clustering was first introduced in our previous work [8]. For the detailed description of the method, refer to [8].

Step 2 - Graph Clustering for a Graphical Representation of Documents

A number of phenomena or systems, such as the Internet [2] have been modeled as networks or graphs. Traditionally those networks were interpreted with Erdos & Rényi's random graph theory, where nodes are randomly distributed and two nodes are connected randomly and uniformly (i.e. Gaussian distribution) [3]. However, researchers have observed that a variety of networks such as those mentioned above, deviate from the random graph theory [1] in that a few most connected nodes are connected to a high fraction of all nodes (there are a few *hub* nodes). However, these *hub* nodes cannot be explained with the traditional random graph theory. Recently, Barabasi and Albert introduced the scale-free network [2]. The scale-free network can explain the *hub* nodes with high degrees because its degree distribution decays as a power law, $P(k) \sim k^{-\gamma}$, where $P(k)$ is the probability that a vertex interacts with k other vertices and γ is the degree exponent [2].

Recently, Ferrer-Cancho and Solé have observed that the graph connecting words in English text follows a scale-free network [4]. Thus, the graphical representation of documents belongs to a highly heterogeneous family of scale-free networks. Our Scale Free Graph Clustering (SFGC) algorithm is based on the scale-free nature (i.e. the existence of a few hub vertices (concepts) in the graphical representation). SFGC starts detecting k hub vertex sets (HVSs) as the centroids of k graph clusters and then assigns the remaining vertices to graph clusters based on the relationships between the remaining objects and k hub vertex sets. For the detailed description of SFGC algorithm, refer to [8].

Step3 - Model-based Document Assignment

In this section, we explain how to assign each document to document clusters. In order to decide which document belongs to which document cluster, CSUGAR matches the graphical representation of each document with each of the graph clusters as models. Here, we might adopt graph similarity mechanisms, such as edit distance (the minimum number of primitive operations for structural modifications on a graph). However, these mechanisms are not appropriate for this task because individual document graphs and graph clusters are too different in terms of the number of vertices and edges. As an alternative to graph similarity mechanisms we take a vote mechanism. This mechanism is based on the classification (HVS or non-HVS) of the vertices in the graph clusters according to their salient scores. This classification leads to different votes. To this end, each vertex of each individual document graph casts two different numbers of votes for document clusters based on whether the vertex belongs to HVS or non-HVS. Each document is assigned to the document cluster that has the majority of votes in the document clusters.

The next three steps correspond to text summarization. Text summarization is to condense information in a set of documents into a concise text. This text summarization problem has been addressed by selecting and ordering sentences in documents based on a salient score mechanism. We address the problem by analyzing the semantic interaction of sentences (as summary elements). This semantic structure of sentences is called Text Semantic Interaction Network (TSIN), where vertices are sentences. We select sentences (vertices in the network) as summary elements based on degree centrality. Unlike traditional approaches, we do not use linguistic features for summarization for MEDLINE abstracts since they usually consist of only single paragraphs.

Step 4 - Making Ontology-enriched Graphical Representations for Each Sentence

The first step of the graphical representation for sentences is basically the same as the graphical representation method for documents except concept extension and individual graph integration. In this step the concepts in sentences are extended using the relationships in relevant document cluster models rather than the entire concept hierarchy. In other words, we extend concepts within relevant semantic field.

Step 5 - Constructing Text Semantic Interaction Network (TSIN)

Text summarization problem has been addressed by selecting and ordering sentences (or phrases) based on various salient score mechanisms. Thus, the key process of text summarization is how to select "salient" sentences (or paragraphs in some approaches) as summary elements. We assume that the sentences becoming summary have the strong semantic relationships with other sentences because summary sentences cover the main points of a set of documents and comprise a condensed version of the set. In order to represent the semantic relationship among sentences, we construct Text Semantic Interaction Network (TSIN), where vertices are sentences, edges are the semantic relationship between them, and edge weights indicate the degree of the relationships.

In order to deal with the semantic relationships between sentences and calculate the similarities (as edge weight in the network) between them, we use edit distance between the graphical representations of sentences. The edit distance between $G1$ and $G2$ is defined as the minimum number of structural modification required to become $G1$ into $G2$, where structural modification is one of vertex insertion, vertex deletion, and vertex update.

Step 6 - Selecting Significant Text Contents for Summary

A number of approaches have been introduced to identify "important" nodes (vertices) in networks (or graphs) for decades. These approaches are normally categorized into degree centrality based approaches and between centrality based approaches. The degree centrality based approaches assume that nodes that have more relationships with others are more likely to be regarded as important in the network because they can directly relate to more other nodes. In other words, the more relationships the nodes in the network have, the more important they are. The betweenness

centrality based approaches views a node as being in a favored position to the extent that the node falls on the geodesic paths between other pairs of nodes in the network [5]. In other words, the more nodes rely on a node to make connections with other nodes, the more important the node is.

These two approaches have their own advantages and disadvantages. Betweenness centrality based approaches yield better experiment results for small graphs to find cluster centroids than other relevant approaches, while they require cubic time complexity so that they are not appropriate for very large graphs. Degree centrality based approaches have been criticized because they only take into account the immediate relationships for each node while they require the linear time complexity and provide comparable output quality with betweenness centrality based approaches.

Because betweenness centrality cannot be applied to very large graphs due to its cubic time complexity, we adopt a well-known hyperlink ranking algorithm, Hypertext Induced Topic Search (HITS) [6], as a centrality measure in a graph. HITS was introduced by Kleinberg in 1998 [6]. HITS algorithm begins with the searching for user's query. The search result, consisting of relevant web pages, is defined as *Root Set*. Then, the *Root Set* is expanded to *Base Set* by adding two kinds of web pages; incoming pages that have hyperlinks to the *Root Set* pages and outgoing pages that are hyperlinked from the *Root Set* pages.

After the input data set is collected, authority and hub scores are calculated for each web page. The authority score of a page is based on the hyperlinks "to" the page while the hub score is based on the links "from" the page. The calculation is based on the following observation:

- If a page has a good authority score, it is meant that many pages that have hyperlinks to the page have good hub scores.
- If a page has a good hub score, the page can give good authority scores to the pages that are hyperlinked by the page.

As they indicate, authority scores mutually reinforce hub scores. Based on these intuitions, for page i , authority score ($A(p_i)$) and hub scores ($H(p_i)$) are mathematically rendered as.

$$A(p_i) = \sum_{p_j \in \{p_j | Link(p_j \rightarrow p_i)\}} H(p_j)$$

$$H(p_i) = \sum_{p_j \in \{p_j | Link(p_i \rightarrow p_j)\}} A(p_j)$$

where, $Link(p_j \rightarrow p_i)$ implies page j (p_j) has a hyperlink to p_i ,

These two iterative operations are performed for each web page; the authority score of each web page is updated with the sum of the hub scores of the web pages that are linked to the page and the hub score of each web page is updated with the sum of the authority scores of the web pages that link to the page. After these two operations are done in each web page, the authority and hub scores are normalized:

$$A(p_i) = \frac{A(p_i)}{\sqrt{\sum_i A(p_i)}} \quad H(p_i) = \frac{H(p_i)}{\sqrt{\sum_i H(p_i)}}$$

However, TSIN graph unlike hyperlinked web is an undirected graph so that we can unify authority score and hub score into

node centrality ($C(N_i)$ for node i), which is mathematically rendered as

$$C(N_i) = \frac{C(N_i)}{\sqrt{\sum_i C(N_i)}}, \quad C(N_i) = \sum_{N_j \in \{N_j | Neighbor(N_j, N_i)\}} C(N_j)$$

where, $Neighbor(N_i, N_j)$ indicates nodes i and j are directly connected each other.

We call this simplified HITS as Mutual Refinement (MR) centrality here since the node centrality is recursively mutually refined. Because the node centrality mutually depends on one another, we provide each node with its degree centrality as an initial value. We will apply MR centrality as well as the degree centrality to measure the centrality of sentences in TSIN.

3. EXPERIMENTAL EVALUATION

In order to measure the effectiveness of CSUGAR, we conducted extensive experiments on public MEDLINE abstracts. For the extensive experiments, first we collected document sets related to various diseases from MEDLINE. We use "MajorTopic" tag along with the disease-related MeSH terms as queries to MEDLINE. After retrieving the base data sets, we generate various document combinations whose numbers of classes are 2 to 9 by randomly mixing the document sets. The document sets used for generating the combinations are later used as answer keys on the document clustering performance measure. For the detailed description about the document sets, the evaluation methods, and the experimental setting, refer to [8].

Because the full detailed experiment results are too big to be depicted in this paper, we average the clustering evaluation metric values and show the standard deviations (σ) for them to indicate how consistent a clustering approach yields document clusters (simply, the reliability of each approach). The σ would be a very important document clustering evaluation factor because document clustering is performed in the circumstance where the information about documents is unknown. Table 1 summarizes the statistical information about clustering results. From the table, we notice the following observations:

- CSUGAR outperforms the nine document clustering methods.
- CSUGAR has the most stable clustering performance regardless of test corpora, while CLUTO Bisecting K-means and K-means do not always show stable clustering performance.
- Hierarchical approaches have a serious scalability problem.
- STC and the original Bisecting K-means have a scalability problem.
- MeSH Ontology improves the clustering solutions of STC.

We observe that CSUGAR has the best performance, yields the most stable clustering results and scales very well. More specifically, CSUGAR shows 45% cluster quality improvement and 72% clustering reliability improvement, in terms of MI, over Bisecting K-means with the best parameters.

Table 1. Summary of Overall Experiment Results on MEDLINE Document Sets

	STC		K-means	Original Bisecting K-means [7]	CLUTO Bisecting K-means		CSUGAR
	word strings	concept strings			Largest	LOS	
MI	μ : 0.429 σ : 0.238	μ : 0.359 σ : 0.149	μ : 0.128 σ : 0.148	μ : 0.395 σ : 0.193	μ : 0.161 σ : 0.139	μ : 0.096 σ : 0.112	μ : 0.053 σ : 0.031
Purity	μ : 0.601 σ : 0.214	μ : 0.731 σ : 0.098	μ : 0.932 σ : 0.080	μ : 0.666 σ : 0.154	μ : 0.918 σ : 0.064	μ : 0.944 σ : 0.056	μ : 0.947 σ : 0.030
F-measure	μ : 0.499 σ : 0.285	μ : 0.512 σ : 0.198	μ : 0.828 σ : 0.206	μ : 0.532 σ : 0.236	μ : 0.780 σ : 0.180	μ : 0.880 σ : 0.139	μ : 0.926 σ : 0.062

LOS: selecting the cluster (to be bisected) with the least overall similarity and Largest: selecting the largest cluster to be bisected. MI: the smaller, the better clustering quality. Purity and F-measure: the bigger, the better clustering quality

Table 2. Experiment Results for Text Summarization: For each document cluster its document cluster model and key sentences as summary are shown

Document Cluster Model for <i>Alzheimer Disease</i>	Top 7 Sentences as Summary for the Document Cluster
	<ul style="list-style-type: none"> • Tau protein extracted from filaments of familial multiple system tauopathy with presenile dementia shows a minor 72-kDa band and two major bands of 64 and 68 kDa that contain mainly hyperphosphorylated four-repeat tau isoforms of 383 and 412 amino acids. • The central pathological cause of Alzheimer disease (AD) is hypothesized to be an excess of beta-amyloid (Abeta) which accumulates into toxic fibrillar deposits within extracellular areas of the brain. These deposits disrupt neural and synaptic function and ultimately lead to neuronal degeneration and dementia • In dementia of Alzheimer type (DAT), cerebral glucose metabolism is reduced in vivo, and enzymes involved in glucose breakdown are impaired in post-mortem brain tissue • Alzheimer's disease (AD), a progressive, degenerative disorder of the brain, is believed to be the most common cause of dementia amongst the elderly • The fundamental cause of Alzheimer dementia is proposed to be Alzheimer disease, i.e. the neurobiological abnormalities in Alzheimer brain • Alzheimer's disease (AD) is a degenerative disease of the brain, and the most common form of dementia • Regional quantitative analysis of NFT in brains of non-demented elderly persons: comparisons with findings in brains of late-onset Alzheimer's disease and limbic NFT dementia.

Document Cluster Model for <i>Osteoarthritis</i>	Top 7 Sentences as Summary for the Document Cluster
	<ul style="list-style-type: none"> • Pathological joint events in both inflammatory arthritis and degenerative arthritis are perpetuated by complex cytokine interactions • In 8, who had severe osteoarthritis, a bicompartamental ICLH (Imperial College-London Hospital) prosthesis was used; in 12, with moderate arthritis, the medial side of the joint was replaced by a unicompartamental Brigham prosthesis • In old scaphoid fractures, the degenerative arthritis begins with an impingement between radial styloid process and proximal pole of the scaphoid, and then reaches the lunocapitate joint. A dorsiflexion by instability is the constant • These patients suffered from painful posttraumatic degenerative arthritis after tarsometatarsal joint fracture-dislocation • More than 85% of all adult cadavers demonstrate degenerative arthritis of the radial subsesamoid joint • OBJECTIVE: Osteoarthritis (OA) is the most common type of arthritis; involvement of joints in the hand is highly prevalent, especially in the elderly • Dysfunction of the pisotriquetral joint: degenerative arthritis treated by excision of the pisiform.

Table 2 shows the experiment results for text summarization for two document clusters (*Alzheimer Disease* and *Osteoarthritis*); due to the page limitation only two document clusters are presented. We believe that its document cluster models in HVS and Top 7 sentences as summary significantly help users understand the document cluster.

4. CONCLUSION

In this paper, we introduce a coherent biomedical literature clustering and summarization approach. This approach takes advantage of ontology-enriched graphical representations of documents. Our approach significantly improves the quality of document clusters and understandability of documents through summaries for each document cluster.

5. ACKNOWLEDGMENTS

This research work is supported in part from the NSF Career grant (NSF IIS 0448023) NSF CCF 0514679 and the PA Dept of Health Tobacco Settlement Formula Grant (#240205, 240196).

6. REFERENCES

- [1] Amaral, L.A.N., Scala, A., Barthélemy, M. and Stanley, H.E. *Proc. Nat. Ac. Sci USA*, 97, 2000, 11149-11152.
- [2] Barabasi, A.L., Albert, R. Emergence of scaling in random networks, *Science*, 286, 1999, 509.
- [3] Erdos, P. and Rényi, A. On the Evolution of Random Graphs. *Publ. Math. Inst. Hungar. Acad. Sci.* 5, 1960, 17-61.
- [4] Ferrer-Cancho, R., and Solé, R.V., The small world of human language. In *Proceedings of the Royal Society of London*, 268, 1482, 2001, 2261–2266.
- [5] Hanneman, R. A., Riddle, M. 2005. Introduction to social network methods [online]. *University of California*. Available from: <http://faculty.ucr.edu/~hanneman/>
- [6] Kleinberg, J., Authorative Sources in a Hyperlinked Environment. In *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*, May 1998.
- [7] Steinbach, M., Karypis, G., and Kumar, V. *A Comparison of Document Clustering Techniques*. Technical Report #00-034. University of Minnesota, 2000.
- [8] Yoo I., Hu X., and Song I.Y., Clustering Ontology-enriched Graph Representation for Biomedical Documents based on Scale-Free Network Theory, accepted in the *IEEE Conference on Intelligent Systems (IEEE IS'06)*, Sept 4-6, 2006.