Biomedical Ontology MeSH Improves Document Clustering Qualify on MEDLINE Articles: A Comparison Study

Illhoi Yoo
College of Information Science and Technology, Drexel University, Philadelphia, PA, 19104
iy28@drexel.edu

Xiaohua Hu
College of Information Science and Technology, Drexel University, Philadelphia, PA, 19104
thu@ischool.drexel.edu

Abstract
Document clustering has been used for better document retrieval, document browsing, and text mining. In this paper, we investigate if biomedical ontology MeSH improves the clustering quality for MEDLINE articles. For this investigation, we perform a comprehensive comparison study of various document clustering approaches such as hierarchical clustering methods (single-link, complete-link, and complete link), Bisecting K-means, K-means, and Suffix Tree Clustering (STC) in terms of efficiency, effectiveness, and scalability. According to our experiment results, biomedical ontology MeSH significantly enhances clustering quality on biomedical documents. In addition, our results show that decent document clustering approaches, such as Bisecting K-means, K-means and STC, gains some benefit from MeSH ontology while hierarchical algorithms showing the poorest clustering quality do not reap the benefit of MeSH ontology.

1. Introduction
Document clustering was initially investigated for improving information retrieval (IR) performance (i.e. precision and recall) because similar documents grouped by document clustering tend to be relevant to the same user queries [7] [9]. However, document clustering has not been widely used in IR systems [3] because document clustering was too slow or infeasible for very large document sets in the early days. As faster clustering algorithms have been introduced, they have been adopted in document clustering. Document clustering has been recently used to facilitate nearest-neighbor search [2], to support an interactive document browsing paradigm [3] [5] [10], and to construct hierarchical topic structures [6]. Thus, as information grows exponentially, document clustering plays a more important role for IR and text mining communities.

2. Background: Ontology and MeSH
An ontology is a formal, explicit specification of a shared conceptualization for a domain of interest [4]. To this end, an ontology is organized by concepts and identifies all the possible relationships among the concepts. Thus, for well-structured ontologies such as Medical Subject Headings (MeSH) (www.nlm.nih.gov/mesh) or Unified Medical Language System (UMLS) (umlsks.nlm.nih.gov), the corresponding domain communities can reach a consensus on the knowledge in the ontologies. For this reason, ontologies can be used as domain knowledge for knowledge-based systems or intelligent agents. We use the MeSH ontology to apply our approach to medical domain.

Medical Subject Headings (MeSH) by the National Library of Medicine 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 only they 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 {“Cancers”, “Tumors”, “Benign Neoplasms”, etc}. MeSH descriptors are organized in a MeSH Tree, which can be seen as a 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. In fact, because descriptors normally appear in more than one place in the tree, they are represented in a graph rather than a tree. In addition to its ontology role, MeSH descriptors were originally used to index MEDLINE articles. For this purpose around 10 to 20 MeSH terms are manually assigned to each article (after reading full papers). On the assignment of MeSH terms to articles around 3 to 5 MeSH terms are set as “MajorTopics” that primarily represent an article.

3. THE USE OF MeSH ONTOLOGY ON VECTOR SPACE MODEL
All document clustering methods are first to convert documents into a proper format. In order to incorporate background knowledge in MeSH ontology into document vector representation, the terms in each document are mapped into MeSH concepts. In order to reduce unnecessary searches rather than searching all Entry terms in each document, 1 to 3-gram words are selected as the candidates of MeSH Entry terms after removing stop words from each document. After matching the candidates with MeSH Entry terms, Entry terms are replaced with Descriptor terms to unite the synonyms or the related terms to descriptors. Then, we eliminate too general MeSH concepts (e.g. HUMAN, WOMEN or MEN) or too common MeSH concepts in MEDLINE articles (e.g. ENGLISH ABSTRACT or DOUBLE-BLIND METHOD). We assume that those terms do not have distinguishable power in clustering documents. The whole process is illustrated in Figure 1. This figure shows that MeSH Entry term sets are detected from “Doc_1” and “Doc_2” documents using the MeSH ontology, and then the Entry terms are replaced with Descriptors based on the MeSH ontology.

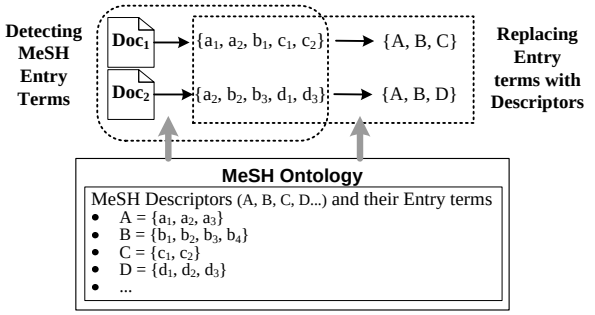

Figure. 1. MeSH Concept Mapping

4. Experimental Evaluation

For test document sets, we first collected document sets related to various diseases from MEDLINE. We use "MajorTopic" tag along with the disease MeSH terms as queries to MEDLINE. After retrieving the data sets, we generate various document combinations whose numbers of classes are 2 to 12 by randomly mixing the document sets. The document sets used for generating the combinations are later used as answer keys on the performance measure. The format of corpus ID in Figure 2 is [C*k*.*n*], where *k* is the number of document sets (classes) and *n* is a sequence number.

We provide all the clustering algorithms except Suffix Tree Clustering (STC) with both word*document matrixes and concept*document matrixes as inputs. For STC, we input both a word string and a concept string (we detected MeSH Entry terms in each string and replaced them with MeSH descriptors). The implementations of STC are based on [10]. We use BiSecting K-means, K-means, and hierarchical clustering algorithms in the CLUTO clustering package[1]. Because BiSecting K-means and K-means may produce different results every time due to their random initializations, we ran them five times.

Figure 2 shows the comparison of MIs for the seven document clustering approaches; Bisecting K-means is classified according to its cluster selection method (LOS: selecting the cluster (to be bisected) with the least overall similarity and Larg.: selecting the largest cluster to be bisected). Table 1 and Table 2 show the overall cluster quality improvement by the use of MeSH ontology on document clustering. From Figure 2 and Table 1 & 2, we notice the following observations.

- MeSH ontology significantly improves clustering solutions on MEDLINE articles for all the document clustering approaches except hierarchical algorithms.
- Hierarchical approaches produce the poorest clustering results and also have the least scalability in our experiment.
- STC gains the maximum benefit from MeSH ontology on MEDLINE document clustering while hierarchical algorithms do not reap the benefit of MeSH ontology.
- STC has a scalability problem; it does not handle document datasets whose sizes are more than 45k.
- Bisecting K-means yields the best clustering solutions on the use of MeSH ontology.

[1] http://www-users.cs.umn.edu/~karypis/cluto/download.html

We observe that biomedical ontology MeSH significantly enhances document clustering quality on MEDLINE articles for the decent approaches (i.e Bisecting K-means, K-means and STC). There are two reasons to support the result. First, the use of ontology on document representation based on vector space model greatly reduces the dimension sizes (i.e. the number of distinct document objects), as shown in Table 3. As Beyer, et al claimed [1], clustering in high dimensional space significantly hampers the similarity detection for documents because distances between every pair of objects tend to the same regardless of data distributions and distance functions. Second, the use of ontology involves only ontology concepts in document representation (i.e. dimension construction) so that simple words not having distinguishable power on clustering documents are eliminated on document representation. As Wang et al pointed out [8], only a small number of words/terms in documents have distinguishable power on clustering documents. Words/terms with distinguishable power are normally the concepts in the domain related to the documents.

5. Conclusion

We perform a fairly comprehensive comparison study of document clustering on 22 MEDLINE corpora for seven document clustering approaches to investigate if biomedical ontology MeSH improves biomedical document clustering quality. Our primary finding is that the decent document clustering approaches (i.e Bisecting K-means, K-means and STC) gains some benefit from MeSH ontology while hierarchical algorithms showing the poorest clustering quality do not reap the benefit of MeSH ontology

6. Reference

[1] Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. When is nearest neighbor meaningful? In *Proc. of 7th Int'l Conference on Database Theory*, 1999, 217-235.

[2] Buckley, C. and Lewit, A. F. Optimization of inverted vector searches. In *Proceedings of SIGIR-85*, 1985, 97-110.

[3] Cutting, D., Karger, D., Pedersen, J. and Tukey, J. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, In *Proceedings of SIGIR '92*, 1992, 318-329.

[4] Hearst, M. A. and Pedersen, J. O. Reexamining the cluster hypothesis: Scatter/Gather on retrieval results. In *Proceedings of SIGIR-96*, 1996, 76–84.

[5] Koller, D. and Sahami, M. Hierarchically classifying documents using very few words. In *Proceedings of ICML-97*, 1997, 170–176.

[6] van Rijsbergen, C. J. *Information Retrieval, 2nd edition*, London: Buttersworth, 1979.

[7] 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.

[8] Willett, P. Recent trends in hierarchical document clustering: A critical review. *Information Processing & Management, 24*, 5, 1988, 577-597.

[9] Zamir, O., and Etzioni O. Web Document Clustering: A Feasibility Demonstration, In *Proceedings of SIGIR 98*, 1998, 46-54.

[10] Gruber, T.R. Towards Principles for the Design of Ontologies used for Knowledge Sharing. *International Journal of Human-Computer Studies, 43*, 1995, 907-928.

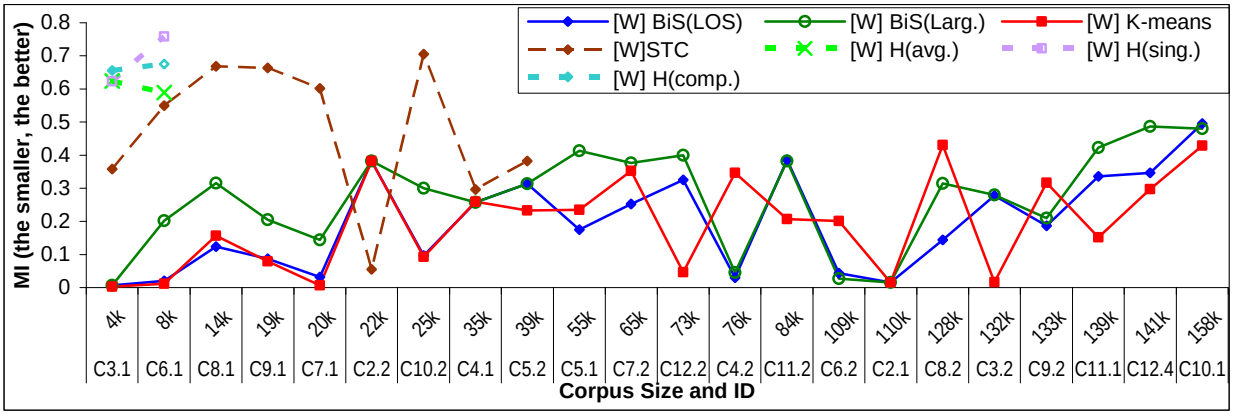

Figure. 2. Comparison of MIs for the Seven Document Clustering Approaches.

Table 1. Overall Cluster Quality Improvement by using Ontology for Bisecting K-means and K-means

	Bisecting K-means (LOS)				Bisecting K-means (Larg.)				K-means			
	Fmeasure	Purity	MI	Entropy	Fmeasure	Purity	MI	Entropy	Fmeasure	Purity	MI	Entropy
Using Words	0.760	0.919	0.197	0.153	0.677	0.891	0.272	0.176	0.761	0.917	0.194	0.141
Using Ontology	0.861	0.957	0.105	0.086	0.686	0.915	0.223	0.124	0.826	0.938	0.115	0.094
Improvement	**13.3%**	**4.1%**	**46.6%**	**44.2%**	**1.3%**	**2.7%**	**18.1%**	**29.3%**	**8.4%**	**2.3%**	**40.9%**	**33.7%**

Note that the smaller MI and Entropy imply the better clustering quality while the bigger F-measure and purity indicate the better clustering quality.

Table 2. Overall Cluster Quality Improvement by using Ontology for Hierarchical Approaches and STC

	Hierarchical (Average-Link)				Hierarchical (Single-Link)				Hierarchical(Complete-Link)				Suffix Tree Clustering			
	F-m.	Pur.	MI	Entr.	F-m.	Pur.	MI	Entr.	F-m.	Pur.	MI	Entr.	F-m.	Pur.	MI	Entr.
Words	0.607	0.161	0.428	0.805	0.691	0.100	0.310	0.985	0.666	0.178	0.400	0.909	0.444	0.554	0.475	0.625
Ontology	0.548	0.257	0.489	0.765	0.690	0.102	0.310	0.985	0.690	0.128	0.337	0.962	0.496	0.742	0.346	0.400
Improv.	-9.7%	59.6%	-14.1%	5.0%	-0.1%	1.5%	-0.2%	0.1%	3.6%	-28.4%	15.8%	-5.9%	**11.8%**	**33.9%**	**27.3%**	**36.1%**

Table 3. Dimension Reduction Rate by the Use of Ontology on Document Representation

Corpus ID	Corpus Size	Concept Vector Dimension Size	Word Vector Dimension Size	Dimension Reduction Rate	Corpus ID	Corpus Size	Concept Vector Dimension Size	Word Vector Dimension Size	Dimension Reduction Rate
C3.1	4k	3080	11243	**72.6%**	C12.2	73k	10575	59987	**82.4%**
C6.1	8k	4809	17804	**73.0%**	C4.2	76k	9859	55068	**82.1%**
C8.1	14k	5955	24454	**75.6%**	C11.2	84k	9930	58972	**83.2%**
C9.1	19k	6818	29337	**76.8%**	C6.2	109k	10983	67760	**83.8%**
C7.1	20k	6521	29066	**77.6%**	C2.1	110k	9229	59936	**84.6%**
C2.2	22k	5718	26252	**78.2%**	C8.2	128k	11499	75757	**84.8%**
C10.2	25k	7534	33532	**77.5%**	C3.2	132k	9716	66450	**85.4%**
C4.1	35k	7567	36379	**79.2%**	C9.2	133k	11819	78360	**84.9%**
C5.2	39k	7961	38894	**79.5%**	C11.1	139k	11010	74333	**85.2%**
C5.1	55k	9521	49208	**80.7%**	C12.4	141k	11973	79998	**85.0%**
C7.2	65k	9883	55129	**82.1%**	C10.1	158k	11117	78086	**85.8%**