

College of Information Science and Technology



Drexel E-Repository and Archive (iDEA)

<http://idea.library.drexel.edu/>

Drexel University Libraries

www.library.drexel.edu

The following item is made available as a courtesy to scholars by the author(s) and Drexel University Library and may contain materials and content, including computer code and tags, artwork, text, graphics, images, and illustrations (Material) which may be protected by copyright law. Unless otherwise noted, the Material is made available for non profit and educational purposes, such as research, teaching and private study. For these limited purposes, you may reproduce (print, download or make copies) the Material without prior permission. All copies must include any copyright notice originally included with the Material. **You must seek permission from the authors or copyright owners for all uses that are not allowed by fair use and other provisions of the U.S. Copyright Law.** The responsibility for making an independent legal assessment and securing any necessary permission rests with persons desiring to reproduce or use the Material.

Please direct questions to archives@drexel.edu

A Comparative Study of Ontology Based Term Similarity Measures on PubMed Document Clustering

Xiaodan Zhang¹, Liping Jing², Xiaohua Hu¹, Michael Ng³, Xiaohua Zhou¹

¹ College of Information Science & Technology, Drexel University, 3141 Chestnut, Philadelphia, PA 19104, USA

² ETI & Department of Math, The University of Hong Kong, Pokfulam Road, Hong Kong.

³ Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong.
{xzhang,thu}@ischool.drexel.edu, lpjing@eti.hku.hk, mng@math.hkbu.edu.hk ,
xiaohua.zhou@drexel.edu

Abstract. Recent research shows that ontology as background knowledge can improve document clustering quality with its concept hierarchy knowledge. Previous studies take term semantic similarity as an important measure to incorporate domain knowledge into clustering process such as clustering initialization and term re-weighting. However, not many studies have been focused on how different types of term similarity measures affect the clustering performance for a certain domain. In this paper, we conduct a comparative study on how different semantic similarity measures of term including path based similarity measure, information content based similarity measure and feature based similarity measure affect document clustering. We evaluate term re-weighting as an important method to integrate domain ontology to clustering process. Meanwhile, we apply k-means clustering on one real-world text dataset, our own corpus generated from PubMed. Experiment results on 8 different semantic measures have shown that: (1) there is no a certain type of similarity measures that significantly outperforms the others; (2) Several similarity measures have rather more stable performance than the others; (3) term re-weighting has positive effects on medical document clustering, but might not be significant when documents are short of terms.

Keywords: Semantic Similarity Measure, Document Clustering, Domain Ontology

1 Introduction

Recent research has been focused on how to integrate domain ontology as background knowledge to document clustering process and shows that ontology can improve document clustering performance with its concept hierarchy knowledge [2, 3, and 16]. Hotho et al. [2] uses WordNet synsets to augment document vector and achieves better results than that of “bag of words” model on public domain. Yoo et al. [16] achieves promising clustering result using MeSH domain ontology for clustering initialization. They first cluster terms by calculating term semantic similarity using MeSH ontology (<http://www.nlm.nih.gov/mesh/>) on PubMed document sets [16]. Then the documents are mapped to the corresponding term cluster. Last, mutual

reinforcement strategy is applied. Varelas et al. [14] uses term re-weighting for information retrieval application. Jing et al. [3] adopt similar technique on document clustering. They re-weight terms and assign more weight to terms that are more semantically similar with each other.

Although existing approaches rely on term semantic similarity measure, not many studies have been done on evaluating the effects of different similarity measures on document clustering for a specific domain. Yoo et al. [16] uses only one similarity measure that calculates the number of shared ancestor concepts and the number of co-occurred documents. Jing et al. [3] compares two ontology based term similarity measure. Even though these approaches are heavily relied on term similarity information and all these similarity measures are domain independent, however, to date, relatively little work has been done on developing and evaluating measures of term similarity for biomedical domain (where there are a growing number of ontologies that organize medical concepts into hierarchies such as MeSH ontology) on document clustering.

Clustering initialization and term re-weighting are two techniques adopted for integrating domain knowledge. In this paper, term re-weighting is chosen because: (1) a document is often full of class-independent “general” terms, how to discount the effect of general terms is a central task. Term re-weighting may help discount the effects of class-independent general terms and aggravate the effects of class-specific “core” terms; (2) hierarchically clustering terms [16] for clustering initialization is more computational expensive and more lack of scalability than that of term re-weighting approach.

As a result, in this paper, we evaluate the effects of different term semantic similarity measures on document clustering using term re-weighting, an important measure for integration domain knowledge. We examine 4 path based similarity measures, 3 information content based similarity measures, and 2 feature based similarity measures for document clustering on PubMed document sets. The rest of the paper is organized as follows: Section 2 describes term semantic similarity measures; section 3 shows document representation and defines the term re-weighting scheme. In section 4, we present and discuss experiment results. Section 5 concludes the paper shortly.

2 Term semantic similarity measure

Ontology based similarity measure has some advantages over other measures. First, ontology is created by human being manually for a domain and thus more precise; second, compared to other methods such as latent semantic indexing, it’s much more computational efficient; Third, it helps integrate domain knowledge into the data mining process. Comparing two terms in a document using ontology information usually exploit the fact that their corresponding concepts within ontology usually have properties in the form of attributes, level of generality or specificity, and their relationships with other concepts [11]. It should be noted that there are many other term semantic similarity measures such as latent semantic indexing, but it’s out of scope of our research, our focus here is on term semantic similarity measure using

ontology information. In the subsequent subsections, we classify the ontology based semantic measures into the following three categories and try to pick popular measures for each category.

2.1 Path based similarity measure

Path based similarity measure usually utilizes the information of the shortest path between two concepts, of the generality or specificity of both concepts in ontology hierarchy, and of their relationships with other concepts.

Wu and Palmer [15] present a similarity measure finding the most specific common concept that subsumes both of the concepts being measured. The path length from most specific shared concept is scaled by the sum of IS-A links from it to the compared two concepts.

$$S_{W\&P}(C_1, C_2) = \frac{2H}{N_1 + N_2 + 2H} \quad (1)$$

In the equation (1), N_1 and N_2 is the number of IS-A links from C_1, C_2 respectively to the most specific common concept C , and H is the number of IS-A links from C to the root of ontology. It scores between 1(for similar concepts) to 0. In practice, we set H to 1 when the parent of the most specific common concept C is the root node.

Li et al. [8] combines the shortest path and the depth of ontology information in a non-linear function:

$$S_{Li}(C_1, C_2) = e^{-\alpha L} \frac{e^{\beta H} - e^{-\beta H}}{e^{\beta H} + e^{-\beta H}} \quad (2)$$

where L stands for the shortest path between two concepts, α and β are parameters scaling the contribution of shortest path length and depth respectively. The value is between 1(for similar concepts) and 0. In our experiment, the same as [8]'s, we set α and β to 0.2 and 0.6 respectively.

Leacock and Chodorow [7] define a similarity measure based on the shortest path $d(C_1, C_2)$ between two concepts and scaling that value by twice the maximum depth of the hierarchy, and then taking the logarithm to smooth the resulting score:

$$S_{L\&C}(C_1, C_2) = -\log(d(C_1, C_2)/2D) \quad (3)$$

where D is the maximum depth of the ontology and similarity value. In practice, we add 1 to both $d(C_1, C_2)$ and $2D$ to avoid $\log(0)$ when the shortest path length is 0.

Mao et al. [10] define a similarity measure using both shortest path information and number of descendents of compared concepts.

$$S_{Mao}(C_1, C_2) = \frac{\delta}{d(C_1, C_2) \log_2(1 + d(C_1) + d(C_2))} \quad (4)$$

where $d(C_1, C_2)$ is the number of edges between C_1 and C_2 , $d(C_1)$ is the number of C_1 's descendants, which represents the generality of the concept. Here, the constant δ refers to a boundary case where C_1 is the only direct hypernym of C_2 , C_2 is the only direct hyponym of C_1 and C_2 has no hyponym. In this case, because the concepts C_1 and C_2 are very close, δ should be chosen close to 1. In practice, we set it to 0.9.

2.2 Information content based measure

Information content based measure associates probabilities with concepts in the ontology. The probability [11] is defined in equation (5), where $freq(C)$ is the frequency of concept C , and $freq(Root)$ is the frequency of root concept of the ontology. In this study, the frequency count assigned to a concept is the sum of the frequency counts of all the terms that map to the concept. Additionally, the frequency counts of every concept includes the frequency counts of subsumed concepts in an IS-A hierarchy.

$$IC(C) = -\log\left(\frac{freq(C)}{freq(Root)}\right) \quad (5)$$

As there may be multiple parents for each concept, two concepts can share parents by multiple paths. We may take the minimum $IC(C)$ when there is more than one shared parents, and then we call concept C the most informative subsumer— $IC_{mis}(C_1, C_2)$. In another word, $IC_{mis}(C_1, C_2)$ has the least probability among all shared subsumer between two concepts.

$$S_{Resnik}(C_1, C_2) = -\log IC_{mis}(C_1, C_2) \quad (6)$$

$$S_{Jiang}(C_1, C_2) = -\log IC(C_1) - \log IC(C_2) + 2\log IC_{mis}(C_1, C_2) \quad (7)$$

Resnik [12] presents a similarity measure. It signifies that the more information two terms share in common, the more similar they are, and the information shared by two terms is indicated by the information content of the term that subsume them in the ontology. It also considers information such as the size of the corpus. Jiang [X] considers not only the shared information between two terms, but also the information content each term contains.

Lin [9] utilizes both the information needed to state the commonality of two terms and the information needed to fully describe these two terms. Since $IC_{mis}(C_1, C_2) \geq \log IC(C_1), \log IC(C_2)$ the similarity value varies between 1 (for similar concepts) and 0.

$$S_{Lin}(C_1, C_2) = \frac{2 \log IC_{mis}(C_1, C_2)}{\log IC(C_1) + \log IC(C_2)} \quad (8)$$

2.3 Feature based measure

Feature based measure assumes that each term is described by a set of terms indicating its properties or features. Then, the more common characteristics two terms have and the less non-common characteristics they have, the more similar the terms are [14]. As there is no describing feature set for MeSH descriptor concepts, in our experimental study, we take all the ancestor nodes of each compared concept as their feature sets. The following measure is defined according to [5, 9]:

$$S_{BasicFeature}(C_1, C_2) = \frac{|Ans(C_1) \cap Ans(C_2)|}{|Ans(C_1) \cup Ans(C_2)|} \quad (9)$$

where $Ans(C_1)$ and $Ans(C_2)$ correspond to description sets (the ancestor nodes) of terms C_1 and C_2 respectively, $C_1 \cap C_2$ is the join of two parent node sets and $C_1 \cup C_2$ is the union of two parent node sets.

Knappe [5] defines a similarity measure as below using the information of generalization and specification of two compared concepts:

$$S_{Knappe}(C_1, C_2) = p \times \frac{|Ans(C_1) \cap Ans(C_2)|}{|Ans(C_1)|} + (1 - p) \times \frac{|Ans(C_1) \cap Ans(C_2)|}{|Ans(C_2)|} \quad (10)$$

where p 's range is $[0, 1]$ that defines the relative importance of generalization vs. specialization. This measure scores between 1 (for similar concepts) and 0. In our experiment, p is set to 0.5.

3 Document representation and re-weighting scheme

MeSH Medical Subject Headings (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 study since they are 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. In addition to its ontology role, MeSH descriptors have been used to index MEDLINE articles. For this purpose, about 10 to 20 MeSH terms are manually assigned to each article (after reading full papers). On the assignment of MeSH terms to articles, about 3 to 5 MeSH terms are set as “MajorTopics” that primarily represent an article.

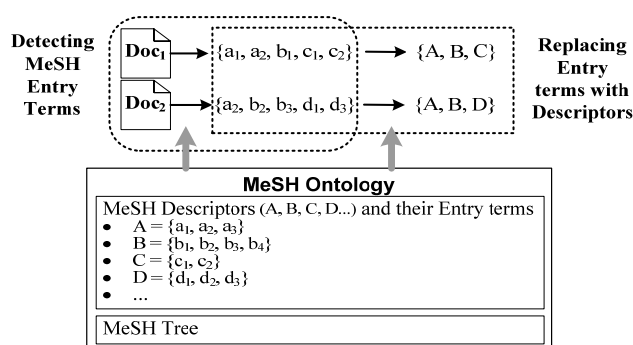


Fig.1. The concept mapping from MeSH entry terms to MeSH descriptors

With mesh descriptor and MeSH tree, the similarity score between two medical terms can be easily calculated. Therefore, we first match the terms in each document abstract to the Entry terms in MeSH and then maps the selected Entry terms into MeSH Descriptors. We select those candidate terms (1- 6gram) that only match with MeSH Entry terms. We then replace those semantically similar Entry terms with the Descriptor term to remove synonyms. We next filter out some MeSH Descriptors that are too general (e.g. HUMAN, WOMEN or MEN) or too common in MEDLINE articles (e.g. ENGLISH ABSTRACT or DOUBLE-BLIND METHOD). We assume that those terms do not have distinguishable power in clustering documents. Hence, we have selected a set of only meaningful corpus-level concepts, in terms of MeSH Descriptors, representing the documents. We call this set *Document Concept Set (DCS)*, where $DCS = \{C_1, C_2, \dots, C_n\}$ and C_i is a corpus-level concept. Fig.1 shows that MeSH Entry term sets are detected from “Doc₁” and “Doc₂” documents using the MeSH ontology, and then the Entry terms are replaced with Descriptors based on the MeSH ontology. For a more comprehensive comparative study, we represent document in two ways: MeSH entry terms, MeSH descriptor terms. At the time of this writing, there are about 23833 unique MeSH descriptor terms, 44978 MeSH ontology nodes (one descriptor term might belong to more than one ontology nodes) and 593626 MeSH entry terms.

Re-weighting Scheme A document is often full of class-independent “general” words and short of class-specific “core” words, which leads to the difficulty of document clustering. Steinbach et al. [13] examines on the data that each class has a “core” vocabulary of words and remaining “general” words may have similar distributions on different classes. To solve this problem, we should “discount” general words and

“emphasize” more importance on core words in a vector [17]. [3, 14] define the term re-weighting scheme as below

$$\tilde{x}_{ji1} = x_{ji1} + \sum_{\substack{i_2=1 \\ i_2 \neq i_1 \\ S(x_{ji1}, x_{ji2}) \geq \text{Threshold}}}^m S(x_{ji1}, x_{ji2}) \cdot x_{ji2} \quad (11)$$

where x stands for term weight, m stands for the number of co-occurred terms, and $S(x_{ji1}, x_{ji2})$ stands for the semantic similarity between two concepts. Through this re-weighting scheme, the weights of semantically similar terms will be co-augmented. Here the threshold stands for minimum similarity score between two compared terms. Since we are only interested in re-weighting those terms that are more semantically similar with each other, it’s necessary to set up a threshold value—the minimum similarity score between compared terms. Besides, it should be noted that the term weight can be referred as term frequency (TF), normalized term frequency (NTF) and TF*IDF (Inverse Document Frequency).

4 Experiment setting and result analysis

4.1 Datasets and indexing schemes

We conduct experiments on public MEDLINE documents (abstracts). First we collect document sets related to various diseases from MEDLINE. We use “MajorTopic” tag along with the disease-related MeSH terms as queries to MEDLINE. Table 1 shows the 10 document sets (24566 documents) retrieved from MEDLINE. Then, the collected dataset is indexed using two schemes: *MeSH entry term* and *MeSH descriptor term*. The average document length for MeSH entry term and MeSH descriptor are 14 and 13 respectively (as shown in table 2). Compared to the average document length—81 when using bag of words representation, the dimension of clustering space is dramatically reduced. A general stop word list is applied to bag of words scheme. Moreover, we collect PubMed documents from 1995-2005 to make MeSH descriptor stop term list for MeSH term and MeSH descriptor term indexing. Since a MeSH entry term can be mapped to more than one MeSH descriptor term in MeSH ontology, we then map it to the MeSH descriptor term which is semantically similar with most of the other terms in the document. For a better comparative study, we also make the following environmental settings: 1) the number of clusters is set to 10, the same as the number of the document sets; 2) documents with length less than 5 are removed from the clustering process; 3) when conducting k-means clustering, we run ten times with random initialization and take the average as the result. During the comparative experiment, each run has the same initialization.

4.2 Evaluation methodology

Cluster quality is evaluated by four extrinsic measures, *entropy* [13], *F-measure* [6], *purity* [18], and *normalized mutual information (NMI)* [1]. Because of space

restrictions, we only describe in detail a recently popular measure—NMI, which is defined as the mutual information between the cluster assignments and a pre-existing labeling of the dataset normalized by the arithmetic mean of the maximum possible entropies of the empirical marginal, i.e.,

$$NMI(X, Y) = \frac{I(X; Y)}{(\log k + \log c) / 2} \quad (12)$$

where X is a random variable for cluster assignments, Y is a random variable for the pre-existing labels on the same data, k is the number of clusters, and c is the number of pre-existing classes. NMI ranges from 0 to 1. The bigger the NMI is the higher quality the clustering is. NMI is better than other common extrinsic measures such as purity and entropy in the sense that it does not necessarily increase when the number of clusters increases. For *Purity and F-measure* ranging from 0 to 1, the bigger the value is the higher quality the clustering has. For entropy, the smaller the value is the higher clustering quality is.

Table 1. The Document Sets and Their Sizes

| | Document Sets | No. of Docs |
|----|----------------------------------|--------------------|
| 1 | Gout | 642 |
| 2 | Chickenpox | 1,083 |
| 3 | Raynaud Disease | 1,153 |
| 4 | Jaundice | 1,486 |
| 5 | Hepatitis B | 1,815 |
| 6 | Hay Fever | 2,632 |
| 7 | Kidney Calculi | 3,071 |
| 8 | Age-related Macular Degeneration | 3,277 |
| 9 | Migraine | 4,174 |
| 10 | Otitis | 5,233 |

Table 2. Document indexing schemes

| Indexing Scheme | No. of term indexed | Avg. doc length |
|------------------------|----------------------------|------------------------|
| MeSH entry term | 14885 | 14 |
| MeSH descriptor term | 8829 | 13 |
| Word | 41208 | 81 |

4.3 Result analysis

To compare the effects of different similarity measures on improving clustering quality, we run k-means clustering on the collected dataset. We represent each document as TF*IDF vector, because this scheme achieves much better performance than NTF and TF. When calculating the distance between one document vector and the cluster center vector, we use cosine similarity measure. Moreover, when representing a document using MeSH entry terms, it's somewhat similar with augmenting a document vector with synonym terms. As one MeSH descriptor term

can relate with many different MeSH entry terms, it is possible that two or more MeSH entry terms with same descriptor term appear in one document. Furthermore, if a document is represented as a document using MeSH descriptors, it can help map all the synonyms occurred in one document to their according descriptor terms. In this paper, we evaluate the clustering qualities of both representation schemes as well as word representation scheme. The process of clustering is as follows: (1) index the document sets using MeSH entry terms or MeSH descriptor terms; (2) calculate term similarity using selected similarity measure and then build similarity matrix for indexed terms; (3) re-weight terms in each document vector using similarity matrix and equation (10); (4) Run k-means clustering. We use dragon toolkit [19] to implement the whole process.

Table 3. Clustering results of MeSH entry terms scheme; each measure is followed by the threshold of similarity value (in parenthesis) that helps achieve the best results.

| Type of Measure | Similarity Measure | Entropy | F-Score | Purity | NMI |
|---------------------|-------------------------|--------------|--------------|--------------|--------------|
| Path based | Wu & Palmer (0.8) | 0.392 | 0.803 | 0.876 | 0.757 |
| | Li et al. (0.7) | 0.353 | 0.830 | 0.871 | 0.771 |
| | Leacock (0.2) | 0.930 | 0.596 | 0.686 | 0.524 |
| | Mao et al. (0.8) | 0.338 | 0.836 | 0.885 | 0.781 |
| Information Content | Resnik (0.0) | 0.353 | 0.821 | 0.877 | 0.774 |
| | Jiang (0.1) | 0.572 | 0.695 | 0.799 | 0.701 |
| | Lin (0.9) | 0.360 | 0.825 | 0.880 | 0.771 |
| Feature based | Basic Feature (0.8) | 0.389 | 0.795 | 0.874 | 0.759 |
| | Knappe (0.8) | 0.484 | 0.778 | 0.831 | 0.717 |
| MeSH entry term | None | 0.353 | 0.800 | 0.870 | 0.774 |
| Word | None | 0.245 | 0.755 | 0.908 | 0.820 |

Experimental results show that of the three types of term similarity measures, there is no a certain type of measures that significantly outperforms others. This can be partially resulted from the fact that most of these measures consider not only the term closeness within the ontology but also the depth of the two compared concepts within the ontology. Apparently, the similarity score of $S_{L\&C}$, S_{Resnik} and S_{Jiang} is not within $[0, 1]$. So term similarity scores using these three measures are normalized before being applied to do term reweighting for a fair comparison reason. Interestingly, Information content based measure with support of corpus statistics has very similar performance with the other two types of measure. This indicates that the corpus statistics is fit with ontology structure of MeSH and does not improve path based measure. The measure of Mao et al. achieves the best result in both indexing schemes as shown in table 3&4. The reason might be that it is the only measure that utilizes the number of descendents information of compared terms. Judging from the overall performance, Wu et al., Li et al., Mao et al., Resink and the two feature based measures have a rather more stable performance than that of others. Moreover, for almost all the cases as shown in table 3, the four evaluation metrics are consistent with each other except that the score of *F-measure* and *Purity* of Wu et al. and Li et al

is slightly better than baseline concept without re-weighting while *NMI* score of them is slightly worse.

Table 4. Clustering results of MeSH descriptor terms scheme; each measure is followed by the threshold of similarity value (in parenthesis) that helps achieve the best results.

| Type of Measure | Similarity Measure | Entropy | F-Score | Purity | NMI |
|---------------------|-------------------------|--------------|--------------|--------------|--------------|
| Path based | Wu & Palmer (0.8) | 0.361 | 0.789 | 0.883 | 0.771 |
| | Li et al. (0.7) | 0.339 | 0.756 | 0.877 | 0.780 |
| | Leacock (0.2) | 0.485 | 0.749 | 0.907 | 0.720 |
| | Mao et al. (0.8) | 0.259 | 0.831 | 0.907 | 0.814 |
| Information Content | Resink (0.0) | 0.346 | 0.815 | 0.890 | 0.777 |
| | Jiang(0.1) | 0.529 | 0.703 | 0.809 | 0.696 |
| | Lin (0.9) | 0.683 | 0.582 | 0.775 | 0.631 |
| Feature based | Basic Feature (0.8) | 0.385 | 0.778 | 0.873 | 0.760 |
| | Knappe (0.8) | 0.375 | 0.784 | 0.866 | 0.765 |
| MeSH descriptor | None | 0.339 | 0.756 | 0.877 | 0.780 |
| Word | None | 0.245 | 0.755 | 0.908 | 0.820 |

From table 3&4, it's easily seen that the overall performance of descriptor scheme is very consistent with and slightly better than that of entry term scheme, which shows that making a document vector more precise by mapping synonym entry terms to one descriptor terms has positive effects on document clustering. It's also noted that both indexing schemes without term re-weighting have competitive performance to those with term re-weighting. It shows that term re-weighting as a method of integrating domain ontology to clustering might not be an effective approach, especially when the documents are short of terms, because when all these terms are very important core terms for the documents, ignoring the effects of some of them by re-weighting can cause serious information loss. This is in contrast to the experiment results in general domain where document length is relatively longer [3].

It's obvious that word indexing scheme achieves the best clustering result although it's not statistically significant (The word scheme experimental result is listed in both table 3&4 for convenience of reader). However, this does not mean indexing medical documents using MeSH entry term or MeSH descriptor is a bad scheme. In other words, it does not mean domain knowledge is not good. First, while keeping competitive clustering results, not only the dimension of clustering space but also the computational cost is dramatically reduced especially when handling large datasets. Second, existing ontologies are under growing, they are still not enough for many text mining applications. For example, there are only 28533 unique entry terms for the time of writing. Third, there is also limitation of term extraction. So far, existing approaches usually use "exact match" to map abstract terms to entry terms and can not judge by the sense the phrase. This will cause serious information loss. For example, when representing document as entry terms, the average document length is 14, while the length of the word representation is 81. Finally, if taking advantage of both medical concept representation and informative word representation, the results of text mining application can be more convincing.

5 Conclusion

In this paper, we evaluate the effects of 9 semantic similarity measures with a term re-weighting method on document clustering of PubMed document sets. The k-means clustering experiment shows that term re-weighting as a method of integrating domain knowledge has some positive effects on medical document clustering, but might not be significant. In detail, we obtain following interesting findings from the experiment by comparing 8 semantic similarity measures three types: path based, information content based and feature based measure with two indexing schemes—MeSH entry term and MeSH descriptor: (1) Descriptor scheme is relatively more effective on clustering than entry term scheme because synonym problem is well handled. (2) There is no a certain type of measures is significantly better than others since most of these measures consider only the path between compared concepts and their depth information within the ontology. (3) Information content based measure using corpus statistics, as well as ontology structure, does not necessarily improve the clustering result when corpus statistics is very consistent with ontology structure (4) As the only similarity measure using the number of descendents information of compared concepts, the measure of Mao et al. has the best clustering result compared to other similarity measure. (5) Similarity measure that is not scoring between 1 and 0 needs to be normalized, otherwise they will aggravate term weight much more aggressively. (6) Over all, term re-weighting achieves similar clustering result with that without term re-weighting. Some of them outperform the baseline, some of them don't and neither of them is very significant, which may indicate that term re-weighting might not be an effective approach when documents are short of terms because when most of these terms are distinguish core terms for a document, ignoring some of them by re-weighting will cause serious information loss. (7) The performance of MeSH term based schemes are slightly worse than that of word based scheme, which can be resulted from the limitation of domain ontology and limitation of term extraction and sense disambiguation. However, while keeping competitive results, indexing using domain ontology dramatically reduces the dimension of clustering space and computational complexity. Furthermore, this finding indicates that there should be an approach taking advantage of both medical concept representation and informative word representation.

In our future work, we may consider other biomedical ontology such as Medical Language System (UMLS) and also expand this comparative study to some public domain.

Acknowledgments. This work is supported in part by NSF Career grant (NSF 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 Banerjee, A. and Ghosh, J. Frequency sensitive competitive learning for clustering on high-dimensional hyperspheres. Proc. IEEE Int. Joint Conference on Neural Networks, pp. 1590-1595.

- 2 Hotho, A., Staab, S. and Stumme, G., "Wordnet improves text document clustering," in Proc. of the Semantic Web Workshop at 26th Annual International ACM SIGIR Conference, Toronto, Canada, 2003.
- 3 Jing, J., Zhou, L., Ng, M. K. and Huang, Z., "Ontology-based distance measure for text clustering," in Proc. of SIAM SDM workshop on text mining, Bethesda, Maryland, USA, 2006.
- 4 Jiang, J.J. and Conrath, D.W., Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In Proceedings of the International Conference on Research in Computational Linguistic, Taiwan, 1998.
- 5 Knappe, R., Bulskov, H. and Andreasen, T.: Perspectives on Ontology-based Querying, International Journal of Intelligent Systems, 2004.
- 6 Larsen, B. and Aone, C. Fast and effective text mining using linear-time document clustering, KDD-99, San Diego, California, 1999, 16-22.
- 7 Leacock, C. and Chodorow, M., Filling in a sparse training space for word sense identification. ms., March 1994.
- 8 Li, Y., Zuhair, A.B., and McLean, D.. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. IEEE Transactions on Knowledge and Data Engineering, 15(4):871-882, July/August 2003.
- 9 Lin, D., Principle-Based Parsing Without Overgeneration. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL'93), pages 112-120, Columbus, Ohio, 1993.
- 10 Mao, W. and Chu, W. W., "Free text medical document retrieval via phrased-based vector space model," in Proc. of AMIA'02, San Antonio, TX, 2002.
- 11 Pedersen, T., Pakhomov, S., Patwardhan, S. and Chute, C., Measures of semantic similarity and relatedness in the biomedical domain. Journal of Biomedical Informatics, In Press, Corrected Proof, June 2006.
- 12 Resnik, O., Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity and Natural Language. *Journal of Artificial Intelligence Research*, 11:95-130, 1999.
- 13 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.
- 14 Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E. G., and Milios, E. E. 2005. Semantic similarity methods in wordNet and their application to information retrieval on the web. WIDM '05. ACM Press, New York, NY, 10-16.
- 15 Wu, Z. and Palmer, M.. Verb Semantics and Lexical Selection. In *Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics (ACL'94)*, pp133-138, Las Cruces, New Mexico, 1994.
- 16 Yoo I., Hu X., Song I-Y., Integration of Semantic-based Bipartite Graph Representation and Mutual Refinement Strategy for Biomedical Literature Clustering, in the Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2006), pp 791-796
- 17 Zhang X., Zhou X., Hu X., Semantic Smoothing for Model-based Document Clustering, accepted in the 2006 IEEE International Conference on Data Mining (ICDM'06).
- 18 Zhao, Y. and Karypis, G. Criterion functions for document clustering: experiments and analysis, Technical Report, Department of Computer Science, University of Minnesota, 2001.
- 19 Zhou, X., Zhang, X., and Hu, X., The Dragon Toolkit, Data Mining & Bioinformatics Lab, iSchool at Drexel University, <http://www.ischool.drexel.edu/dmbio/dragontool>