Medical Document Clustering Using Ontology-Based Term Similarity Measures

Xiaodan Zhang, Drexel University, USA
Liping Jing, The University of Hong Kong, China
Xiaohua Hu, Drexel University, USA
Michael Ng, Hong Kong Baptist University, China
Jiali Xia, Jiangxi University of Finance and Economics, China
Xiaohua Zhou, Drexel University, USA

ABSTRACT

150 words or less

Keywords: document clustering; domain ontology; semantic similarity measure

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 (Hothe et. al., 2003; Jing et. al., 2006; Yoo et. al., 2006). Hothe, Staab and Stumme (2003) employed WordNet synsets to augment document vector and achieves bet-
ter results than that of “bag of words” model on public domain. Yoo, Hu, and Song (2006) applied MeSH domain ontology to clustering initialization and achieved promising clustering results. Terms are first clustered by calculating semantic similarity using MeSH ontology (http://www.nlm.nih.gov/mesh/) on PubMed document sets. Then the documents are mapped to the corresponding term cluster. Last, mutual reinforcement strategy is applied. Varelas et al. (2005) integrated domain ontology using term re-weighting for information retrieval application. Terms are assigned more weight if they are semantically similar with each other. Jing et al. (2006) adopted similar technique on document clustering.

Although existing approaches rely on term semantic similarity, not many studies have been done on evaluating the effects of different similarity measures on document clustering for a specific domain. Yoo, Hu, and Song (2006) employed one similarity measure that calculates the number of shared ancestor concepts and the number of co-occurred documents. Jing et al. (2006) compared 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 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. In our previous study (Zhang et al., 2007), a comparative study is conducted on a selected PubMed document set. However, the conclusion on one dataset may not be very general. Moreover, the similarity score threshold applied in previous study brings unfairness to term re-weighting since the distribution of similarity scores are different in terms of different similarity measure. Therefore, for a fair comparison, we use the minimum path length between two documents as the threshold.

Clustering initialization and term re-weighting are two techniques adopted for integrating domain knowledge. In this article, 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 is more possible to help discount the effects of class-independent general terms and thus aggravates the effects of class-specific “core” terms; (2) hierarchically clustering terms (Yoo, Hu, & Song, 2006) for clustering initialization is more computational, expensive and more lack of scalability than that of term re-weighting approach.

As a result, 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 four path-based similarity measures, three information content-based similarity measures, and two feature-based similarity measures for document clustering on two biomedical literature sets: Disease10 and OHSUMED23. The rest of the article is organized as follows: the “Term Semantic Similarity Measures” section describes term semantic similarity measures; the “Methodology” section shows document representation and defines the term re-weighting scheme. The “Datasets” section list two biomedical data sets. In the “Experimental Results And Analysis” section, we present and discuss experiment results. The last section briefly concludes the article.

TERM SEMANTIC SIMILARITY MEASURES

Ontology-based similarity measure has some advantages over other measures. First, ontology is manually created by human beings for a domain and thus more precise; second, compared to other methods such as latent semantic indexing, it is 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 exploits 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 (Pedersen et al., 2007). It is worth noting that there are many other term semantic similarity measures such as latent semantic indexing, but this is out of scope of this study and our focus 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.

**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 (1994) developed 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.

\[
S_{W&P}(C_1, C_2) = \frac{2H}{N_1 + N_2 + 2H}
\]

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. The similarity score is between 1 (for similar concepts) and 0. In our experiment, we set \(a\) and \(b\) to 0.2 and 0.6, respectively for the best performance.

Leacock and Chodorow (1994) defined 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:

\[
S_{L&C}(C_1, C_2) = -\log \left( \frac{d(C_1, C_2)}{2D} \right)
\]

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 and Chu (2002) presented a similarity measure using both shortest path information and number of descendents of compared concepts.

\[
S_{Mao}(C_1, C_2) = \frac{\delta}{d(C_1, C_2) \log_2(1 + d(C_1) + d(C_2))}
\]

where \(d(C_1, C_2)\) is the number of edges between \(C_1\) and \(C_2\), \(d(C_r)\) is the number of \(C_r\)'s descendents, which represents the generality of the concept. Here, the constant \(\delta\) 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, \(\delta\) should be chosen close to 1. In practice, we set it to 0.9.

**Information Content-Based Measure**

Information content-based measure associates probabilities with concepts in the ontology. The probability 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 (Pedersen et al., 2007). In this study, the frequency count assigned to a concept is the sum of the frequency counts of all the concepts and 0. In our experiment, we set \(a\) and \(b\) to 0.2 and 0.6, respectively for the best performance.
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.

\[ IC(C) = -\log \left( \frac{\text{freq}(C)}{\text{freq}(\text{Root})} \right) \]  

(5)

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

\[ S_{\text{Resnik}}(C_1, C_2) = -\log IC_{\text{mis}}(C_1, C_2) \]  

(6)

\[ S_{\text{Jiang}}(C_1, C_2) = -\log IC(C_1) - \log IC(C_2) + 2 \log IC_{\text{mis}}(C_1, C_2) \]  

(7)

Resnik (1999) defined a similarity measure that 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. The measure reveals information about the usage within corpus of the part of the ontology queried. Jiang and Conrath (1998) included not only the shared information content between two terms, but also the information content each term contains.

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

\[ S_{\text{Lin}}(C_1, C_2) = \frac{2 \log IC_{\text{mis}}(C_1, C_2)}{\log IC(C_1) + \log IC(C_2)} \]  

(8)

**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 (Varelas et al., 2005). 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 based on the discussion in Knappe et al. (2006) and Lin (1003):

\[ S_{\text{BasicFeature}}(C_1, C_2) = \frac{|\text{Ans}(C_1) \cap \text{Ans}(C_2)|}{|\text{Ans}(C_1) \cup \text{Ans}(C_2)|} \]  

(9)

where \( \text{Ans}(C_i) \) and \( \text{Ans}(C_j) \) correspond to description sets (the ancestor nodes) of terms \( C_i \) and \( C_j \) respectively, \( C_i \cap C_j \) is the join of two parent node sets and \( C_i \cup C_j \) is the union of two parent node sets.

Knappe et al. (2006) developed a similarity measure, as seen below, using the information of generalization and specification of two compared concepts:

\[ S_{\text{Knappe}}(C_1, C_2) = p \times \frac{|\text{Ans}(C_1) \cap \text{Ans}(C_2)|}{|\text{Ans}(C_1)|} + (1 - p) \times \frac{|\text{Ans}(C_1) \cap \text{Ans}(C_2)|}{|\text{Ans}(C_2)|} \]  

(10)

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

**Methodology**

Given a document set, our clustering method is composed of the following steps: (1) apply ontology to index whole document set; each document is thus represented as a vector of terms; (2) each term’s weight is re-calculated by the proposed term re-weighting method; (3) Spherical K-means is run the on the dataset.
Mesh Ontology

Ontology is very important to biomedical documents clustering. First, biomedical literature is usually composed of many complicated biomedical concepts with name variations containing usually more than one word. Second, bag-of-words model suffers from “the curse of dimension” and lacks interpretation power to clustering results. Therefore, MeSH ontology is applied to index biomedical literature in this article.

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. 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”}. As a result, descriptors terms are used in this research. 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 PubMed 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 three to five MeSH terms are set as “Major-Topic” that primarily represent the article. This indicates that submitting Major Topic MeSH term query to PubMeD usually retrieves dataset with ground truth.

Mesh Concept Indexing

Terms in each document are mapped to the entry terms in MeSH and then maps the selected Entry terms into MeSH Descriptors to remove synonyms.

In detail, our indexing system matches the terms in each document to the entry terms in MeSH and then maps the selected entry terms into MeSH Descriptors. Instead of searching all entry terms in the MeSH against each document, we select 1- to 5-gram words as the candidates of MeSH Entry terms. Then, only those candidate terms are chosen that match with MeSH entry terms. We then replace those semantically similar entry terms with the Descriptor term to remove synonyms. Next, some MeSH Descriptors are filtered out that are too common or have nothing do with the contents of PubMed articles (e.g., ENGLISH ABSTRACT or Government supported); A stop term list is generated for this purpose by analyzing 10 years of PubMed documents (1994-2004). At the time of this writing, there are about 23,833 unique MeSH descriptor terms, 44,978 MeSH

Figure 1. PubMed document indexing

<table>
<thead>
<tr>
<th>MeSH Ontology</th>
<th>MeSH Descriptors (A, B, C, D, ...) and their Entry terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>A = {a1, a2, a3}</td>
<td>B = {b1, b2, b3, b4}</td>
</tr>
<tr>
<td>C = {c1, c2}</td>
<td>D = {d1, d2, d3}</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td>MeSH Tree</td>
<td></td>
</tr>
</tbody>
</table>
ontology nodes (one descriptor term might belong to more than one ontology nodes) and 593,626 MeSH entry terms.

MeSH entry term sets are detected from documents “Doc1” and “Doc2,” using the MeSH ontology, and then the Entry terms are replaced with Descriptors based on the MeSH ontology. Both MeSH descriptors and entry terms are multi-grams.

**Term Re-Weighting**

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. (2000) examined 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. Jiang and Conrath (1998) and Varelas et al. (2005) define the term re-weighting scheme as below,

$$\tilde{x}_{jl1} = x_{jl1} + \sum_{j_2 \neq j_1}^{m} S(x_{jl1}, x_{jl2}) x_{jl2}$$

where $x$ stands for term weight, $m$ stands for the number of co-occurred terms, and $S(x_{jl1}, x_{jl2})$ stands for the similarity score between two concepts. Through this re-weighting scheme, the weights of semantically similar terms will be co-augmented. Since we are only interested in re-weighting those terms that are more semantically similar with each other, it is necessary to set up a threshold value—the minimum similarity score or the minimum path length between compared concepts. In practice, document is first represented according to certain scheme such as TF-IDF. Then, each term’s weight is augmented by equation (Pedersen, Pakhomov, Patwardhan, & Chutte, 2007).

**DATASETS**

**Disease10**

Disease10 dataset is collected from PubMed (a Web interface of Medline documents) by submitting queries using “MajorTopic” tag along with the corresponding MeSH term of the disease name. For example, if the disease name’s corresponding MeSH term is “Gout,” then the query will become “Gout [Major Topic].” Table 1 shows the ten classes of document sets and their document numbers (24,566 documents). The document class name is the query name. The average document length for MeSH descriptor is 13 (as shown in Table 2). Compared to the average document length—81—when using

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

Table 1. The document sets and their sizes
bag of words representation, the dimension of clustering space is dramatically reduced. A general stop term list is applied to bag of words scheme.

**OHSUMED23**

OHSUMED consists of scientific abstracts collected from Medline, an online medical information database. The selected OHSUMED corpus contains 13,929 Medline abstracts of the year 1991, each of which was assigned with one or multiple labels out of 23 cardiovascular diseases categories. Excluding abstracts with multiple labels, we indexed the rest 7,400 abstracts belonging to 23 classes.

**EXPERIMENTAL RESULTS AND ANALYSIS**

**Evaluation Methodology**

Cluster quality is evaluated by four extrinsic measures, *entropy* (Steinbach, Karypis & Kumar, 2000), *F-measure* (Larsen & Aone, 1999), *purity* (Zhao & Karypis, 2001), and *normalized mutual information* (NMI) (Banerjee & Ghosh, 2002). NMI 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, that is,

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

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. Purity can be interpreted as the classification rate under the assumption that all samples of a cluster are predicted to be members of the actual dominant class for that cluster. Entropy is a more comprehensive measure than purity since rather than just considering the number of objects “in” and “not in” the most frequent class, it considers the entire distribution. F-score combines the information of precision and recall which is the extensively applied in information retrieval, with values falling in \([0, 1]\) and the larger is the value, the better is the cluster quality.

**Experiment Settings**

To improve the efficiency of the calculation of term-term similarity, a 44,978 term-term similarity matrix (including all MeSH descriptors) is constructed for each similarity measure before the document vector re-weighting.

The similarity score is disregarded between two terms whose minimal path length larger than 3, since we are only interested augmenting the weights of terms that are similar enough. This is better than setting a similarity score threshold and very important to evaluate different semantic similarity measures in a fair manner. The distributions of the similarity scores between documents are usually various in terms of different similarity measures. Setting one score threshold to all similarity measures can make the results easily biased toward several measures and need time consuming tuning (Zhang et al., 2007). Therefore, we apply minimum length threshold instead of similarity score threshold.

**Table 2. Document indexing schemes**

<table>
<thead>
<tr>
<th>Indexing Scheme</th>
<th>No. of term indexed</th>
<th>Avg. doc length</th>
</tr>
</thead>
<tbody>
<tr>
<td>MeSH descriptor term</td>
<td>8829</td>
<td>13</td>
</tr>
<tr>
<td>Word</td>
<td>41,208</td>
<td>81</td>
</tr>
</tbody>
</table>

Copyright © 2008, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
The minimum path length is defined as:

\[
\text{MinLen}(C_1, C_2) = \text{Dep}(C_1) + \text{Dep}(C_2) - 2 \cdot \text{Dep}(C_1, C_2)
\]

where \(\text{Dep}(C_i)\) indicates the depth of concept \(C_i\) within the ontology and \(\text{Dep}(C_1, C_2)\) is the depth of the nearest co-parent of concept \(C_1\), \(C_2\).

 Apparently, the similarity score range of \(S_{L\&C}^\text{Resink}\) and \(S_{\text{Jiang}}^\text{Resink}\) is not within \([0, 1]\). For a fair comparison, their similarity matrices are normalized before they are applied to re-weighting document vector. In detail, each similarity score is denominated by the row sum. In this study, each document is represented as TF-IDF vector since this scheme achieves much better performance than normalized term frequency and term frequency (Zhang, Zhou, & Hu, 2006). Each document vector is re-weighted using equation the equation by Pedersen, Pakhomov, Patwardhan, and Chute (2007) and the ontology term-term similarity matrix. Spherical K-means is used for documents clustering, for it is a well-known vector-based clustering algorithm. Documents are also indexed using unigram words for a more comprehensive comparison. Documents are not considered in our experiments if they contain fewer than five terms. The whole process is implemented using dragon toolkit (Zhou, Zhang, & Hu, 2006).

**Result Analysis**

Table 3 and 4 show the experimental results of document clustering on Disease10 and OHSUMED23 datasets, respectively. The nine ontology-based similarity measures are divided by their corresponding types including: path-based, information-content-based and feature-based. “MeSH descriptor” and “Word” indicate the type of document representation and they do not use term re-weighting scheme.

**Comparison Between “Re-Weighting” and “None Re-Weighting”**

The performance between re-weighting and none re-weighting varies in terms of the corresponding datasets. For Disease10 dataset, most similarity measures slightly outperform none re-weighting, that is, MeSH descriptor. For OHSUMED23 dataset, the results of different schemes are very close. Three measures including Li, Zuhair and McLean (2003), Leacock and Chodorow (1994) and Resnik (1999) have slightly better performances than none re-weighting scheme. These results show that the re-weighting scheme can slightly improve document clustering, but it is not very signifi-

<table>
<thead>
<tr>
<th>Type of Measure</th>
<th>Measure Name</th>
<th>Entropy</th>
<th>F-Score</th>
<th>Purity</th>
<th>NMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Path-based</td>
<td>Wu &amp; Palmer</td>
<td>0.348</td>
<td>0.858</td>
<td>0.874</td>
<td>0.779</td>
</tr>
<tr>
<td></td>
<td>Li et al.</td>
<td>0.304</td>
<td>0.834</td>
<td>0.901</td>
<td>0.799</td>
</tr>
<tr>
<td></td>
<td>Leacock</td>
<td>0.276</td>
<td>0.853</td>
<td><strong>0.923</strong></td>
<td><strong>0.811</strong></td>
</tr>
<tr>
<td></td>
<td>Mao et al.</td>
<td>0.342</td>
<td>0.830</td>
<td>0.875</td>
<td>0.782</td>
</tr>
<tr>
<td>Information-Content-based</td>
<td>Resink</td>
<td>0.295</td>
<td>0.856</td>
<td>0.906</td>
<td>0.802</td>
</tr>
<tr>
<td></td>
<td>Jiang</td>
<td>0.300</td>
<td>0.848</td>
<td>0.905</td>
<td>0.800</td>
</tr>
<tr>
<td></td>
<td>Lin</td>
<td>0.342</td>
<td>0.845</td>
<td>0.882</td>
<td>0.782</td>
</tr>
<tr>
<td>Feature-based</td>
<td>Basic Feature</td>
<td>0.358</td>
<td>0.818</td>
<td>0.872</td>
<td>0.775</td>
</tr>
<tr>
<td></td>
<td>Knappe</td>
<td>0.350</td>
<td>0.834</td>
<td>0.876</td>
<td>0.778</td>
</tr>
<tr>
<td></td>
<td>MeSH descriptor</td>
<td>0.341</td>
<td>0.772</td>
<td>0.867</td>
<td>0.776</td>
</tr>
<tr>
<td></td>
<td>Word</td>
<td>0.245</td>
<td>0.755</td>
<td>0.908</td>
<td>0.820</td>
</tr>
</tbody>
</table>

Table 3. Clustering results of Disease10

Copyright © 2008, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
They also show that term re-weighting as a method of integrating domain ontology to clustering might not be a very effective approach when the documents are short of terms (Table 2), 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 on the contrary to the experiment results (Jing et al., 2006) in general domain where document length is relatively longer.

### Comparison Between Different Similarity Measures

Experimental results on two datasets show that, among the three types of term similarity measures, there is no certain type of measure that significantly outperforms others. Interestingly, information-content-based measures, with the support of corpus statistics, have very similar performances with the other two types of measure. This may indicate that the corpus statistics is fit with ontology structure of MeSH and thus does not have better performance than path-based measures. Two path-based measures including Leacock and Chodorow (1994) and Li, Zuhair, and McLean (2003) achieve the best performance on both datasets, respectively. Both measures consider not only the shortest path and depth of two concepts. Judging from the overall performance on the two datasets, Li, Zuhair, and McLean (2003), Leacock and Chodorow (1994), Mao and Chu (2002), Resnik (1999) and Jiang and Conrath (1998) have rather more stable performances than that of the other measures. Feature-based measures always have the worst performance. This shows that using parent concepts as features may have negative impact on term re-weighting.

### Comparison Between Ontology-Based and Word-Based Document Representation

The performance of word scheme is significantly different on the two datasets. For Disease10 dataset, word scheme is slightly better than ontology-based scheme, but this is not significant. On OHSUMED23 dataset, word scheme performs significantly worse than the other schemes. The results show both advantage of ontology and the limitation of ontology. First, while keeping competitive or significantly better clustering results, not only the dimension of clustering space but also the computational cost are dramatically reduced especially when handling large datasets. Second, existing ontologies are under growing, they are still not enough for many text mining applications. For

<table>
<thead>
<tr>
<th>Type of Measure</th>
<th>Measure Name</th>
<th>Entropy</th>
<th>F-Score</th>
<th>Purity</th>
<th>NMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Path-based</td>
<td>Wu &amp; Palmer</td>
<td>2.209</td>
<td>0.244</td>
<td>0.347</td>
<td>0.165</td>
</tr>
<tr>
<td></td>
<td>Li et al.</td>
<td>2.181</td>
<td>0.253</td>
<td>0.356</td>
<td>0.174</td>
</tr>
<tr>
<td></td>
<td>Leacock</td>
<td>2.199</td>
<td>0.241</td>
<td>0.351</td>
<td>0.168</td>
</tr>
<tr>
<td></td>
<td>Mao et al.</td>
<td>2.183</td>
<td>0.255</td>
<td>0.354</td>
<td>0.173</td>
</tr>
<tr>
<td>Information-Content-based</td>
<td>Resnik</td>
<td>2.194</td>
<td>0.252</td>
<td>0.352</td>
<td>0.170</td>
</tr>
<tr>
<td></td>
<td>Jiang</td>
<td>2.199</td>
<td>0.251</td>
<td>0.351</td>
<td>0.168</td>
</tr>
<tr>
<td></td>
<td>Lin</td>
<td>2.234</td>
<td>0.239</td>
<td>0.341</td>
<td>0.158</td>
</tr>
<tr>
<td>Feature-based</td>
<td>Basic Feature</td>
<td>2.219</td>
<td>0.241</td>
<td>0.344</td>
<td>0.162</td>
</tr>
<tr>
<td></td>
<td>Knappe</td>
<td>2.226</td>
<td>0.239</td>
<td>0.340</td>
<td>0.160</td>
</tr>
<tr>
<td></td>
<td>MeSH descriptor</td>
<td>2.193</td>
<td>0.248</td>
<td>0.353</td>
<td>0.170</td>
</tr>
<tr>
<td></td>
<td>Word</td>
<td>2.321</td>
<td>0.200</td>
<td>0.302</td>
<td>0.130</td>
</tr>
</tbody>
</table>

Table 4. Clustering results of OHSUMED23

Copyright © 2008, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
example, there are only about 44,000 unique MeSH descriptor 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. This will cause serious information loss. For example, when representing document as MeSH descriptor terms, the average document length is only 13 for Disease10, while the length of the corresponding 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.

CONCLUSION AND FUTURE WORK
In this article, we evaluate the effects of nine semantic similarity measures with a term re-weighting method on document clustering of PubMed document sets. The spherical k-means clustering experiment shows that term re-weighting has some positive effects on medical document clustering, but might not be very significant. In detail, we obtain following meaningful findings by comparing nine semantic similarity measures three types: path-based, information-content-based and feature-based measure with two indexing schemes—MeSH descriptor and Word: (1) term re-weighting achieves very similar clustering results with none term re-weighting. Some of them outperform none re-weighting, some of them do not and neither of them is very significant, which indicates that term re-weighting can be effective in a very limited degree when documents are short of terms because when most of these terms are distinguishable core terms for a document, ignoring some of them by re-weighting will cause information loss; more developed ontology and advanced term extraction technique may help term re-weighting achieve better results; (2) There is no certain type of measure that is significantly better than others; the best performance are achieved by two path-based measures including Leacock and Chodorow (1994) and Li, Zuhair, and McLean (2003) that consider both the closeness and the depth of the compared concepts; feature-based measures have the worst overall performance, which shows that using parent concepts as feature set is not effective for this application; although information-content-based measures consider both ontology and corpus statistics, they do not achieve better results than the other measure types; (3) the performance of MeSH scheme is much better than that of word scheme on OHSUMED23 dataset and slightly worse than word scheme on Disease10 dataset, which demonstrates both the advantage and limitation of domain ontology; while keeping competitive or significantly better results, indexing using MeSH ontology dramatically reduces the dimension of clustering space and computational complexity; however, the limitation of ontology such as limited concepts and rough term extraction techniques can cause information loss easily and thus hurt the clustering performance. 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.

ACKNOWLEDGMENT
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


Xiaodan Zhang is a PhD student in College of Information Science and Technology, Drexel University. He received his BS in library and information science from Northeast Normal University, China in 1997 and MS in computer science from Jinan University, China in 2003. His current research interests include graph-based, ontology-based and model-based text data mining.

Liping Jing is a PhD student in the Department of Mathematics and E-Business Technology Institute at the University of Hong Kong. She received her Bachelor of Science and Master of Philosophy degrees in computer science from the Northern Jiaotong University, in 2000 and 2003, respectively. Her research interests are in the areas of data mining, subspace clustering algorithm, ontology-based programming and business intelligence.

Xiaohua Hu is currently an associate professor and the founding director of the Data Mining and Bioinformatics Laboratory in the College of Information Science and Technology at Drexel University. His current research interest includes biomedical literature data mining, bioinformatics, text mining, semantic Web mining and reasoning, rough set theory and application, information extraction, and information retrieval. He has published more than 140 peer reviewed research papers in various journals, conferences, and books, and co-edited nine books/proceedings. He has received a few prestigious awards, including the 2005 NSF Faculty Early Career Development (NSF Career) Award, the Best Paper Award from the 2007 International Conference on Artificial Intelligence, the Best Paper Award from the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, the 2006 IEEE Granular Computing Outstanding Service Award, and the 2001 IEEE Data Mining Outstanding Service Award. He is the founding editor-in-chief of the International Journal of Data Mining and Bioinformatics, and associate editor/editorial board member of four international journals. His research projects are funded by the US National Science Foundation (NSF), US Department of Education, and the Pennsylvania Department of Health.

Michael Ng is a professor in the Department of Mathematics at the Hong Kong Baptist University. As an applied mathematician, Michael’s main research areas include bioinformatics, data mining, operations research and scientific computing. Michael has published and edited 5 books, published more than 160 journal papers. He is the principal editor of the Journal of Computational and Applied Mathematics, and the associate editor of SIAM Journal on Scientific Computing. http://www.math.hkbu.edu.hk/~mng/

Jiali Xia is a full professor and dean of the UFSoft School of Software, Jiangxi University of Finance & Economics. She obtained her BSc, MSc and PhD degrees all from the Computer Science Department of Huazhong University of Science and Technology in 1987, 1993, and 2003, respectively. Her research interest is real-time embedded database system. She has published many papers on top Chinese journals such as China Computer Transaction and China Software Transaction. She is the second-class prize winner of the Science and Technology Advancement Award from the Chinese Ministry of Education, in 2004 and third-class winner of the Science and Technology Advancement Award from Jiaxi Province in 2007.

Xiaohua Zhou is a PhD candidate in College of Information Science and Technology, Drexel University. He received his bachelor’s and master’s degree from Shanghai Jiao Tong University, China in 1999 and 2002, respectively. His current research interests include information retrieval, text data mining, and information extraction.