Chapter 7.27
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

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 article, we conduct a comparative study on how different term semantic similarity measures including path-based, information-content-based and feature-based similarity measure affect document clustering. Term re-weighting of document vector is an important method to integrate domain ontology to clustering process. In detail, the weight of a term is augmented by the weights of its co-occurred concepts. Spherical k-means are used for evaluate document vector re-weighting on two real-world datasets: Disease10 and OHSUMED23. Experimental results on nine different semantic measures have shown that: (1) there is no 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.

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 (Hotho et. al., 2003; Jing et. al., 2006; Yoo et. al., 2006). Hotho, Staab and Stumme (2003) employed WordNet synsets to augment document vector and achieves better 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 cluttering 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 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 pervious 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 lists two biomedical data sets. In the "Experimental Results And Analysis" section, we present and discuss experiment results. The last section briefly concludes the article.