Expert Systems with Applications 42 (2015) 2264–2275

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

# A semantic approach for text clustering using WordNet and lexical chains

Tingting Wei [a], Yonghe Lu [c,*], Huiyou Chang [b], Qiang Zhou [a], Xianyu Bao [d]

[a] Department of Information Science and Technology, Sun Yat-Sen University, Guangzhou, China
[b] Department of Software, Sun Yat-Sen University, Guangzhou, China
[c] Department of Information Management, Sun Yat-Sen University, Guangzhou, China
[d] Shenzhen Academy of Inspection and Quarantine, Shenzhen, China

ARTICLE INFO

*Article history:*
Available online 18 October 2014

*Keywords:*
Text clustering
WordNet
Lexical chains
Core semantic features

ABSTRACT

Traditional clustering algorithms do not consider the semantic relationships among words so that cannot accurately represent the meaning of documents. To overcome this problem, introducing semantic information from ontology such as WordNet has been widely used to improve the quality of text clustering. However, there still exist several challenges, such as synonym and polysemy, high dimensionality, extracting core semantics from texts, and assigning appropriate description for the generated clusters. In this paper, we report our attempt towards integrating WordNet with lexical chains to alleviate these problems. The proposed approach exploits ontology hierarchical structure and relations to provide a more accurate assessment of the similarity between terms for word sense disambiguation. Furthermore, we introduce lexical chains to extract a set of semantically related words from texts, which can represent the semantic content of the texts. Although lexical chains have been extensively used in text summarization, their potential impact on text clustering problem has not been fully investigated. Our integrated way can identify the theme of documents based on the disambiguated core features extracted, and in parallel downsize the dimensions of feature space. The experimental results using the proposed framework on reuters-21578 show that clustering performance improves significantly compared to several classical methods.



## 1. Introduction

Text clustering is a useful technique that aims at organizing large document collections into smaller meaningful and manageable groups, which plays an important role in information retrieval, browsing and comprehension. Traditional clustering algorithms are usually relying on the BOW (Bag of Words) approach, and an obvious disadvantage of the BOW is that it ignores the semantic relationship among words so that cannot accurately represent the meaning of documents. As the rapid growth of text documents, the textual data have become diversity of vocabulary, they are high-dimensional, and carry also semantic information. Therefore, text clustering techniques that can correctly represent the theme of documents and improve clustering performance, ideally process data with a small size, are greatly needed. Recently, a number of semantic-based approaches are being developed. WordNet (Miller, 1995), which is one of the most widely used thesauruses for English, has been extensively used to improve the quality of text clustering with its semantic relations of terms (Amine, Elberrichi, & Simonet, 2010; Bouras & Tsogkas, 2012; Chen, Tseng, & Liang, 2010; Dang, Zhang, Lu, & Zhang, 2013; Fodeh, Punch, & Tan, 2011; Hotho, Staab, & Stumme, 2003; Jing, Zhou, Ng, & Huang, 2006; Kang, Kim, & Lee, 2005; Recupero, 2007; Sedding & Kazakov, 2004; Song, Li, & Park, 2009).

However, there still exist several challenges for the clustering results. (1) Synonym and polysemy problems. There has been much work done on the use of ontology to replace the original terms in a document by the most appropriate ontology concept for the solution of these problems; this process is known as word sense disambiguation (WSD). This approach, however, has not proven to be as useful as first hoped. For example, approaches that expand the feature space by replacing a term with its potential concepts only increase the feature space without necessarily improving clustering performance (Fodeh et al., 2011; Hotho et al., 2003). (2) High-dimensional term features. High dimension

* Corresponding author.
*E-mail addresses:* tingtingwei2011@126.com (T. Wei), luyonghe@mail.sysu.edu.cn (Y. Lu), isschy@mail.sysu.edu.cn (H. Chang), gfs_007@163.com (Q. Zhou), baoxianyu@163.com (X. Bao).

http://dx.doi.org/10.1016/j.eswa.2014.10.023
0957-4174/

of feature space may increase the processing time and diminish the clustering performance, which is a key problem in text clustering. Most current techniques usually rely on matrix operation methods such as LSI (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990), ICA (Hyvärinen & Oja, 2000), and LDA (Martínez & Kak, 2001) to deal with this problem. Unfortunately, these models need too much computation. Although there also exist a few techniques have considered semantic information (Recupero, 2007; Termier, Sebag, & Rousset, 2001), they have various weaknesses. For example, they do not explicitly and systematically consider the theme of a document. (3) Extract core semantics from texts. Existing dimension-reduced methods may remove some topic features, which results in the semantic content of a document is decomposed and cannot be reflected. It is desirable to extract a subset of the disambiguated terms with their relations (known as the core semantic features) that are "cluster-aware", which leads to improving the clustering accuracy with a reduced number of terms. (4) Assign distinguished and meaningful description for the generated clusters. In order to conveniently recognize the content of each cluster, it is necessarily to assign concise and descriptive labels to help analysts to interpret the result. Nevertheless, good solutions of assigning topic labels to clusters for ease of analysis, recognition, and interpretation are still rare.

This paper attempts to alleviate mentioned above problems, its contributions can be summarized as follows.

(1) We propose a modified similarity measure based on WordNet for word sense disambiguation. This is based on the idea that the explicit and implicit semantic relationships between synsets (concepts) in WordNet impose equally importance factors in the word similarity measure. Previous works have showed that exploiting the structural information of WordNet can improve the accuracy of similarity measurement, but the effects of adding textual data to structural information are still not very extensively researched. In this paper, we explore if the combination of the structural information and the glosses of synsets can provide a more accurate assessment of the similarity between terms for word sense disambiguation.
(2) We introduce lexical chains to capture the main theme of texts. Although lexical chains have been extensively used in text summarization, their potential impact on text clustering problem has not been fully investigated. In our work, we investigate the identification of lexical chains for text representation. We have observed that the concepts extracted from lexical chains as a small subset of the semantic features can ideally cover the theme of texts, and sufficient to reduce the dimensions of feature set, potentially leading to better clustering results.
(3) We show that our method can estimate the true number of clusters by observing the experimental results, which is valuable for determining the value of $k$ in K-means clustering algorithms.
(4) We also demonstrate that our generated topical labels have good indicator of recognizing and understanding the content of clusters. Since lexical chains represent the most of semantic content of texts, we believe that topic labels which describe and interpret the content of a cluster should be selected from the words of lexical chains. In this paper, we use the disambiguated concepts (word senses) from lexical chains in the selection of topic labels for the generated clusters, this solution is especially important when the concept title is ambiguous.

The rest of the paper is organized as follows: Section 2 reviews some related works. Section 3 presents a modified similarity measure based on WordNet for word sense disambiguation. In Section 4, we describe how to extract core semantics by using lexical chains. Section 5 details the experiments that evaluate our method and the analysis of results. Finally, we conclude this work and show its implications in Section 6.

## 2. Related works

To date, text clustering has been heavily researched and a huge variety of techniques has been proposed to deal with it. The goal of the clustering process is to group the documents which are similar in contents into a single group. In order to understand our work better, some relevant works about several research fields related to our interests will be introduced and the limitations of the described approaches will be presented as well.

### 2.1. WordNet

WordNet is one of the most widely used and largest lexical databases of English. In general as a dictionary, WordNet covers some specific terms from every subject related to their terms. It maps all the stemmed words from the standard documents into their specifies lexical categories. In this approach the WordNet 2.1 is used which contains 155,327 terms, 117,597 senses, and 207,016 pairs of term-sense. It groups nouns, verbs, adjectives and adverbs into sets of synonyms called synsets. The synsets are organized into senses, giving thus the synonyms of each word, and also into hyponym/hypernym (i.e. Is-A), and meronym/holonym (i.e. Part-Of) relationships, providing a hierarchical tree-like structure for each term.

The application of incorporating semantic features from the WordNet (Miller & Charles, 1991) lexical database has been widely used to improve the accuracy of text clustering techniques. For example, Dave et al. (Dave, Lawrence, & Pennock, 2003) employed synsets as features for document representation and subsequent clustering. However, word sense disambiguation was not performed, and WordNet synsets actually decreased clustering performance. Accordingly, Hotho et al. (2003) used WordNet in document clustering for word sense disambiguation to improve the clustering performance. Sedding and Kazakov (2004) extended this work by exploring the benefits of disambiguating the terms using their part of speech tags. The main limitation of both approaches is the increase in dimensionality of the data. Gharib, Fouad, and Aref (2010) matched the stemmed keywords to concepts in WordNet for word sense disambiguation. Their approach improves the efficiency of the applied clustering algorithms; however, it seems to over generalize the affected keywords (Bouras & Tsogkas, 2012). In the study of Amine et al. (2010), the authors accepted that the assignment of terms to concepts in ontology can be ambiguous and can lead to loss of information in their attempt to reduce dimensionality.

### 2.2. Semantic similarity

Semantic similarity plays an important role in natural language processing, information retrieval, text summarization, text categorization, text clustering and so on. In recent years the measures based on WordNet have attracted great concern. Many semantic similarity measures have been proposed. In general, all the measures can be grouped into four classes: path length based measures, information content based measures, feature based measures, and hybrid measures. An exhaustive overview of these approaches can be found in (Meng, Huang, & Gu, 2013). Following the cited overview, we focus on measures that are related to our work.