Data & Knowledge Engineering xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

Editorial

A fuzzy document clustering approach based on domain-specified ontology

Lin Yue [a,b,c], Wanli Zuo [a,b,*], Tao Peng [a,b], Ying Wang [a,b], Xuming Han [d]

[a] College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, PR China
[b] Key Laboratory of Symbolic Computation and Knowledge Engineering of the Ministry of Education, Changchun, Jilin 130012, PR China
[c] School of Information Technology and Electrical Engineering, the University of Queensland, Brisbane 4072, Australia
[d] School of Computer Science and Engineering, Changchun University of Technology, Jilin 130012, PR China

ARTICLE INFO

Article history:
Received 26 November 2013
Received in revised form 10 April 2015
Accepted 25 April 2015
Available online xxxx

Keywords:
Domain-specified ontology
Document clustering
Feature selection
Singular value decomposition (SVD)
Fuzzy equivalence relation

ABSTRACT

Document clustering techniques include automatic document organization, topic extraction, fast information retrieval or filtering, etc. Numerous methods have been developed for document clustering research. Despite the advances achieved, however, document clustering still presents certain challenges such as optimizing feature selection for low-dimensional document representation and incorporating mutual information between the documents into a clustering algorithm. This paper mainly focuses on these two questions. First, we construct a domain-specific ontology that provides the controlled vocabulary describing the hazards related to dairy products. Synonyms of the controlled vocabulary in document set are considered to be relatively prevalent and fundamentally important for feature selection. Second, in combination with the vector space model (*VSM*), we perform singular value decomposition (*SVD*) to translate all of the term-document vectors into a concept space. We then obtain the mutual information between documents by calculating the similarity of every two document vectors in the orthogonal matrix of right singular vectors. As the mutual information matrix is also a fuzzy compatible relation, a fuzzy equivalence can be derived by calculating max–min transitive closure. Finally, based on the fuzzy equivalence relation, all of the data sequences are easily allocated into clusters under the guidance of a cluster validation index. Our method both reduces the dimensionality of the original data and considers the correlation between the terms. The experimental results show that encoding the ontologies in the aggregation process could provide better clustering results. Moreover, the proposed work has been applied to food safety supervision which is beneficial for government and society.

1. Introduction

Document clustering (or text clustering) is the application of cluster analysis to textual documents [1]. It has applications in automatic document organization, topic extraction, fast information retrieval or filtering, etc. Document clustering was initially investigated for improving the precision or recall in information retrieval systems by allocating documents into previously unseen categories. Driven by the sheer size and dimensionality of data, the focus has shifted towards providing ways to browse collections of documents or to organize the search results for display usually in a structured or hierarchical manner [2–5].

The rapidly growing availability of unstructured textual data, such as blog postings, emails, online review forums or discussion board messages, has brought a need to improve document clustering especially in different specified-domains. Because the documents from source articles for clustering may have been written by different groups, from different viewpoints, or have different

[*] Corresponding author at: College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, PR China.
E-mail address: wanli@jlu.edu.cn (W. Zuo).

http://dx.doi.org/10.1016/j.datak.2015.04.008
0169-023X/© 2015 Elsevier B.V. All rights reserved.

writing style, clustering these textual materials is therefore a challenge due to the diversity of vocabulary used and the general lack of guidance in terms of background knowledge that could provide domain information. Clustering with the high dimensionality of document representation usually causes poor performance of clustering results and affects the efficiency of clustering algorithms. Thus, optimizing the means of feature selection for low-dimensional document representation is very important in the document-clustering task.

As a comprehensive semantic lexical ontology for the English language, WordNet [6] has been widely used for document clustering. With the guidance of lexical ontology, clustering results can be improved by optimizing the means of feature representation. The existing work can be classified into two types, namely, the concept mapping method and the embedded method. In the concept mapping method, either the terms in the document set are replaced by a concept extracted from ontology or the related concepts are added into the term vector for document representation [7–12]. In short, this type of method has not proven to be as effective as hoped because it variously increases the dimensionality of the data or decreases information of the raw dataset. In the embedded method, the correlation between terms is considered [13–15]. Because embedded methods do not explicitly identify the semantic features of a document, they cannot be used to interpret the resulting clusters. Furthermore, this method does not reduce the dimensionality of the original data.

Because domain ontology is a formal explicit specification of a shared conceptualization for a domain of interest, more and more domain-specified ontologies have been constructed by domain experts; these represent specific domain knowledge of various clustering methods for improving performance [16–19]. Generally, a common vocabulary is usually defined by researchers who need to share information in a domain. An example is gene ontology (*GO*) [20], which provides a controlled vocabulary to describe gene and gene product attributes in any organism and which has resulted in numerous methods being proposed to exploit biomedical literature [2–4,21]. Medical Subject Headings (MeSH ontology) [22], as another example, is a huge controlled vocabulary that organizes medical terms into a hierarchical structure for the purpose of indexing; it has been used to facilitate the text mining process [23]. These two well-known biomedical ontologies have been widely used to exploit the conceptual structure of biomedical document collections. In addition, the National Natural Science Foundation of China (*NSFC*) receives a large number of research proposals every year, which need to be grouped according to their similarities in research disciplines and assigned to appropriate experts for peer review. Previous methods for grouping these proposals are based on manual matching of similar research discipline areas or keywords. The problem is that the exact research discipline areas of the proposals cannot often be accurately designated by the applicants due to their subjective views and possible misinterpretations. Thus, research ontology (*RO*) has been constructed for research project selection, which along with rich information in the proposals' full text can be used to cluster the research proposals effectively [24]. Constructing this type of domain-specific ontology is beneficial for both governments and society. Analogously, our project is concerned with economic and livelihood issues (i.e., mining food complaint information). Specifically, let us suppose that several different websites contain dairy product complaint information or provide food safety supervision services. If these websites share and publish the same underlying ontology of the terms they all use, then computer agents can extract and aggregate information from these different sites. The agents can send this aggregated information to food safety agency or to relevant enterprises as warning messages. Because complaint documents contain considerable domain-specific terminology, they are difficult to represent with appropriate and low-dimensional features.

As the development of Semantic Web technologies enables us to structure domain knowledge into an ontology, in our paper, we have constructed a dairy ontology (*DO*)with Protégé [25], which provides a controlled vocabulary to describe dairy and hazards associated with dairy products. We have then analyzed dairy complaint documents based on this domain ontology. This ontology is built according to Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880 [26]. To overcome the limitation of one domain ontology, HowNet [27], a lexical ontology for the Chinese language, is used to incorporate background knowledge for identifying the synonymous terms of the controlled vocabulary in document sets. The approach mainly focuses on identification of domain-specific terminologies; this procedure simply provides initial feature terms for document representation. Then, in combination with the vector space model, we perform the singular value decomposition (*SVD*) to translate all of the term-document vectors into a concept space. After that, we obtain the mutual information between documents by calculating the similarity of every two document vectors in the orthogonal matrix of right singular vectors. The mutual information matrix is also a fuzzy compatible relation from which a fuzzy equivalence is derived by calculating max–min transitive closure. Finally, based on the fuzzy equivalence relation, all of the data sequences are easily allocated into clusters under the guidance of a cluster validation index. It is a contribution because the validation index we proposed can determine the best partitions by considering a low inter-cluster relation and a high intra-cluster relation. Moreover, our method both reduces the dimensionality of the original data and considers the correlation between terms, which could be treated as a combination of the concept mapping and embedded methods for document clustering. Among others, our method is suitable for providing an effective solution to the main challenges in document clustering which are gigantic volume, high dimensionality and complex semantics.

The rest of this paper is organized as follows. Section 2 reviews the related work. In Section 3, we first describe feature selection methodology with the aid of lexical ontology HowNet and domain-specified ontology. Second, we briefly review the mathematical theories of fuzzy sets theory used in our paper. In Section 4, we introduce the fuzzy clustering method with details. Finally, we will focus on how to determine the best number of clusters. Section 5 presents the experimental results and Section 6 concludes the paper.

2. Related work

Historically, ontologies arise out of the branch of philosophy that addresses the nature of reality and is concerned with analyzing various types or modes of existence, often with special attention to the relations between essence and existence, between intrinsic