Data & Knowledge Engineering 69 (2010) 1208–1226

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

Editorial

An integration of WordNet and fuzzy association rule mining for multi-label document clustering

Chun-Ling Chen [a], Frank S.C. Tseng [b,*], Tyne Liang [a]

[a] Department of Computer Science, National Chiao Tung University, HsinChu 300, Taiwan, ROC
[b] Dept. of Information Management, National Kaohsiung 1st University of Science & Technology, YanChao, Kaohsiung 824, Taiwan, ROC

ARTICLE INFO

Available online 25 September 2010

Keywords:
Fuzzy association rule mining
Text mining
Document clustering
WordNet
Frequent itemsets

ABSTRACT

With the rapid growth of text documents, document clustering has become one of the main techniques for organizing large amount of documents into a small number of meaningful clusters. However, there still exist several challenges for document clustering, such as high dimensionality, scalability, accuracy, meaningful cluster labels, overlapping clusters, and extracting semantics from texts. In order to improve the quality of document clustering results, we propose an effective Fuzzy-based Multi-label Document Clustering (FMDC) approach that integrates fuzzy association rule mining with an existing ontology WordNet to alleviate these problems. In our approach, the key terms will be extracted from the document set, and the initial representation of all documents is further enriched by using hypernyms of WordNet in order to exploit the semantic relations between terms. Then, a fuzzy association rule mining algorithm for texts is employed to discover a set of highly-related fuzzy frequent itemsets, which contain key terms to be regarded as the labels of the candidate clusters. Finally, each document is dispatched into more than one target cluster by referring to these candidate clusters, and then the highly similar target clusters are merged. We conducted experiments to evaluate the performance based on Classic, Re0, R8, and WebKB datasets. The experimental results proved that our approach outperforms the influential document clustering methods with higher accuracy. Therefore, our approach not only provides more general and meaningful labels for documents, but also effectively generates overlapping clusters.

1. Introduction

The incessant flourishing of Internet invigorates various textual documents to be shared over the cyberspace astonishingly. However, it also makes users suffer from the information-overloading problem. In particular, when users pose queries to WWW search engines, they usually bewilderingly receive a small number of relevant Web pages intermingled with a large number of irrelevant Web pages.

To effectively manage and organize the result of a search engine query, there inspires the study of document clustering techniques. The aim of this study is to automatically discover the hidden similarity and the key concepts of clustered documents for users to comprehend a large amount of documents. Over the past decades, several effective document clustering algorithms have been proposed to mitigate the hassle, including the *k*-means [1], Bisecting *k*-means [2], Hierarchical Agglomerative Clustering

* Corresponding author. Present/permanent address: 1, University Road, YanChao, Kaoshiung County, Taiwan 824, ROC.
E-mail addresses: chunling@cs.nctu.edu.tw (C.-L. Chen), imfrank@ccms.nkfust.edu.tw (F.S.C. Tseng), tliang@cs.nctu.edu.tw (T. Liang).

0169-023X/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2010.08.003

(HAC) [3], and Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [4]. Nevertheless, as pointed out by [5–9], there are still challenges in improving the clustering quality, which we list as follows:

- *To cope with high dimensionality*: as the volume of textual document increases, the dimensionality of term features increases as well.
- *To improve the scalability*: many document clustering algorithms work fine on small document sets, but fail to deal with large document sets efficiently.
- *To promote the accuracy*: many existing document clustering algorithms require users to specify the number of clusters as an input parameter. However, it is difficult to determine the number of clusters in advance. Moreover, an incorrect estimation of the input parameter, i.e., the number of clusters, may lead to poor clustering accuracy [6].
- *To assign meaningful cluster labels*: meaningful cluster labels will guide users in the process of browsing the retrieved results. Thus, each cluster should be labeled with an understandable description. However, most of the traditional clustering algorithms do not provide labels for clusters.
- *To enable overlapping clusters*: many well-known clustering algorithms focus on hard clustering, where each document belongs to exactly one cluster. However, a document could contain multiple subjects. By using soft clustering algorithms [9], a document would appear in multiple clusters (i.e., overlapping clusters).
- *To extract semantics from text*: the bag-of-words representation used for clustering algorithms is often unsatisfactory as it ignores the conceptual similarity of terms that do not co-occur actually [5,7].

To resolve the problems of high dimensionality, large size, and understandable cluster description, Beil et al. [8] developed the first frequent itemsets-based algorithm, namely Hierarchical Frequent Term-based Clustering (HFTC), where the frequent itemsets are generated based on the association rule mining [10]. They only considered the low-dimensional frequent itemsets as clusters. Moreover, HFTC discovers overlapping clusters, which is useful for a search engine where overlapping clusters occur like Yahoo! Directory.

However, the experiments of Fung et al. [6] showed that HFTC is not scalable. For a scalable algorithm, Fung et al. proposed the FIHC (Frequent Itemset-based Hierarchical Clustering) algorithm by using frequent itemsets derived from association rule mining to construct a hierarchical topic tree for clusters. They also proved that using frequent itemsets for document clustering can reduce the dimensionality of term features effectively. Yu et al. [11] presented another frequent itemset-based algorithm, called TDC, to improve the clustering quality and scalability. This algorithm dynamically generates a topic directory from a document set using only closed frequent itemsets and further reduces the dimensionality. But, the clusters generated by FIHC and TDC are non-overlapping. In [12], the authors proposed that document clustering methods should provide multiple subjective perspectives onto the same document to enhance their practical applicability.

Recently, WordNet [13], one of the most widely adopted thesaurus for English, has been extensively used as an ontology in grouping documents with its semantic relations of terms [5,7,14,15]. Many existing document clustering algorithms mainly transform text documents into simplistic flat bags of document representation, i.e., term vectors or bags of keywords. Once terms are treated as individual items in such simplistic representation, the semantic content of a document is decomposed and cannot be reflected. Thus, Dave et al. [14] proposed using synsets as features for document representation and subsequent clustering. However, synsets decrease the clustering performance in all experiments without considering word sense disambiguation. Meanwhile, Hotho et al. [5] used WordNet in document clustering for word sense disambiguation to improve the clustering results. Jing et al. [15] presented another application of WordNet, which described how to find mutual information between terms by using the background knowledge through WordNet. In [7], Recupero proposed a new unsupervised document clustering method by using WordNet lexical and conceptual relations to allow common clustering algorithms to perform well. In this paper, the reasons of utilizing hypernyms from WordNet are two-fold:

(1) We intend to obtain more general and conceptual labels for derived clusters.
(2) From the experimental results in [14,16], the authors found that the performance of adding hypernyms is better than adding synonymy.

Among the techniques developed for data and text mining, association rule mining [10] is one of the useful and successful techniques for discovering interesting rules. It helps users discover meaningful association rules to represent a relationship between different pairs of a set of attribute values. The form of an association rule can be represented as $X \rightarrow Y$, where X and Y are sets of items and $X \cap Y = \emptyset$, and is usually adopted for market basket analysis to describe the following meaning: customers that buy product X also buy product Y for satisfying some predefined *minimum support value* and *minimum confidence value*. In general, each itemset has an associated measure of statistical significance called *Support* value, which is the fraction of all transactions that contain the itemset. For example, an itemset X with support value, $\mathrm{supp}(X) = 0.5$, regards there are 50% of transactions in the dataset containing X. An itemset can be chosen as a *frequent itemset* if its support value is larger than or equal to the predefined *minimum support value*. The *confidence* value of an association rule, denoted $\mathrm{conf}(X \rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X)$, is to measure how often items in Y appear in transactions which also contain X. Finally, a rule $X \rightarrow Y$ will be discovered whether its confidence value is larger than or equal to the predefined *minimum confidence value* or not.

However, there are still two situations to be confronted, if we use association rule mining in our approach:

(1) Some important terms that express the topics of a document may be rarely appeared in the document collection. That is, only the terms which frequently occur in the document collection can be obtained, which implies the important sparse terms may be obscured in the process of document clustering.