Pattern Recognition 000 (2017) 1–13

Contents lists available at ScienceDirect

Pattern Recognition

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

Concept decompositions for short text clustering by identifying word communities

Caiyan Jia[a,*], Matthew B. Carson[b], Xiaoyang Wang[a], Jian Yu[a]

[a] School of Computer and Information Technology & Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China
[b] Division of Health and Biomedical Informatics, Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA

ARTICLE INFO

Article history:
Received 11 July 2016
Revised 23 August 2017
Accepted 30 September 2017
Available online xxx

MSC:
00-01
99-00

Keywords:
Short text clustering
Concept decomposition
Spherical *k*-means
Semantic word community
Community detection

ABSTRACT

Short text clustering is an increasingly important methodology but faces the challenges of sparsity and high-dimensionality of text data. Previous concept decomposition methods have obtained concept vectors via the centroids of clusters using *k*-means-type clustering algorithms on normal, full texts. In this study, we propose a new concept decomposition method that creates concept vectors by identifying semantic word communities from a weighted word co-occurrence network extracted from a short text corpus or a subset thereof. The cluster memberships of short texts are then estimated by mapping the original short texts to the learned semantic concept vectors. The proposed method is not only robust to the sparsity of short text corpora but also overcomes the curse of dimensionality, scaling to a large number of short text inputs due to the concept vectors being obtained from term-term instead of document-term space. Experimental tests have shown that the proposed method outperforms state-of-the-art algorithms.

1. Introduction

In the current Web 2.0 era, an increasing number of short texts have been generated including search result snippets, forum titles, image or video titles and tags, frequently asked questions, tweets, microblogs, and so on. This has resulted in a growing need for fast and efficient clustering of short texts according to high similarities within and dissimilarities between clusters. A well-designed short text clustering algorithm has the ability to greatly stimulate and promote its real applications such as topic detection, answering service recommendations, image or video tagging, information retrieval, etc. However, unlike normal texts, use of short texts is complicated by sparsity and high dimensionality. As a result, the classical *tf-idf* (term frequency-inverse document frequency) measure, the vector space model (VSM), and normal text clustering methods may not work well when applied to short texts.

One way to solve the sparsity issue of feature vectors is to expand short texts to long texts by the use of external knowledge sources such as Wikipedia [1], WordNet [2], HowNet [3], Web search results [4], other user constructed knowledge bases [5–7], peripheral information sources [8,9], etc. However, these external knowledge-enhanced methods have the following two issues. First, the creation and maintenance of such resources (e.g., Wikipedia and WordNet) can be very expensive. Second, this introduces the new challenge of how to properly use those external resources (see [10] as an example). In most cases, solving this new problem itself is time-consuming and complicated.

Alternatively, some research efforts have concentrated on improving traditional methods of normal text clustering or designing new models to handle short texts. Existing methods include probabilistic topic models (BTM [11] and GSDMM [12]), and heuristic optimization methods (nonnegative matrix factorization methods [13,14], extended vector space models [15,16], and other heuristic and search snippet specific methods [17–19]). The main idea of these methods such as BTM, TNMF [14], and Generalized VSM [15] is to make use of the relationships between pairs of terms in order to compensate for the sparsity of short texts. However, these methods ignore the relationships among three or more terms. As evidenced by the psychologist's statement "Concepts are the glue that holds our mental world together" [20], words related to the same topic are likely to co-occur in the same text, thus they link together and form densely connected communities in the word co-occurrence network of a corpus. Therefore, it is possible to ex-

[*] Corresponding author.
E-mail address: cyjia@bjtu.edu.cn (C. Jia).

https://doi.org/10.1016/j.patcog.2017.09.045
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

tract concept vectors directly from the word co-occurrence network rather than from the document-term space (the classical method of the latter is spherical k-means [21]). The effect can be seen in methods such as conceptual grouping [22] and word sense induction [23], both of which concentrate on fine-grained word clusters and have not been successfully used to cluster short texts to the best of our knowledge.

In addition, short texts are very sparse, thus their terms are particularly valuable. For example, corpora with even millions of short texts may only contain a few thousand terms that characterize them. Accordingly, extracting concept vectors from term-term space is beneficial for overcoming the curse of dimensionality for large-scale corpora. Therefore, inspired by the concept decomposition method spherical k-means [21] in normal text clustering, we propose a novel concept decomposition method, WordCom, which is based on the identification of semantic word communities using a k-means-type community detection method.

The procedure of WordCom has four steps. First, we construct the word co-occurrence network for a corpus. Second, we extract the semantic word communities from the network using k-means-type algorithm k-rank-D [24]. In k-rank-D, the initial cluster centers and the number of clusters are determined by actively selecting k potential centers located in the right upper part of the decision graph, which characterizes the likelihood of data points being cluster centers by a higher density than their neighbors and by a relatively large distance from points with higher densities [25]. Third, we combine the word communities and their corresponding centers to form concept vectors. Finally, we project all short texts into these concept vectors and obtain their cluster memberships. Moreover, for a small subset of a large-scale short text corpus, words in the subset may have already covered most of words in the whole corpus. This allows us to obtain the concept centers of the corpus only from its subset, which makes our proposed concept decomposition method scale easily for very large short text corpora.

The remainder of this paper is organized as follows. Section 2 introduces the background of this study including related studies and the basic idea of concept decomposition. Section 3 presents our new proposed concept decomposition method, WordCom. Section 4 shows the comparison results. Section 5 draws conclusions and further considerations.

2. Background

2.1. From normal text to short text clustering

Probabilistic topic models such as PLSA (Probabilistic Latent Semantic Analysis) [26] and LDA (Latent Dirichlet Allocation) [27] are classical methods for uncovering hidden topics from normal text corpora. In these models, both term distribution within topics and topic distribution in texts can be inferred by maximum likelihood estimation methods. To address the sparseness of short texts using the LDA topic model, the BTM (biterm topic model) has been developed to capture term co-occurrence pattern implied in short texts by extending the traditional unigram LDA model to 2-gram LDA model [11]. Subsequently, another variation of LDA, GSDMM (a collapsed Gibbs sampling algorithm of Dirichlet Multinomial Mixture), was proposed to cluster short text corpora and also showed good performance on normal text clustering [12].

Similar to topic models, non-negative matrix factorization (NMF) has the ability to identify hidden structures of terms and texts on topics represented by two matrix factors: term-topic matrix **U** and topic-document matrix **V** [28]. Ncut (a similarity weighted NMF) was developed to tackle the sparsity issue and cluster short text corpora [13]. TNMF (a two-step NMF framework) [14], was proposed later. This method performs symmetric NFM on a term similarity matrix to define a term-topic matrix **U**, followed by inference of the topic-document matrix **V** using the NMF framework on the original term-document matrix X according to the learned **U**.

The vector space model (VSM) is a classical model for representing normal texts. VSM assumes that terms are independent and it ignores the semantic relationships among terms. Therefore, in order to overcome the sparseness of terms in a short text corpus, a generalized VSM [15] is used to represent short texts, where the correlations between pairs of terms are used rather than the weights of single terms. Similarly, the method in [16] uses the group of related keywords extracted from the processed short texts themselves to expand the short text corpus and alleviate its sparsity.

TermCut is a core term-based bisect clustering method [17] for short text clustering. It finds the best term by optimizing the clustering criterion RMcut at each step of bisection. Therefore, at each bisection, it must traverse all remaining terms to find the 'core term'. The high dimensionality of short texts results in a high level of time complexity for TermCut.

2.2. Concept decomposition and spherical k-means

Dhillon et al. [21] have proposed a concept decomposition method named spherical k-means to cluster normal texts. Concept decomposition was later extended to concept factorization [29–31]. In this section, we will introduce the basic idea of concept decomposition.

Let $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_n}\}$ be the text vectors of a corpus. Each text is represented by a vector ($\mathbf{x_i}$, $i = 1, \dots, n$) of m possible terms x_{ij}, $j = 1, \dots, m$, where each element of the vector tracks the weight of a term in the text. Usually the weight is measured by the *tf-idf* value, i.e., $x_{ij} = tf_{ij} \times \log(\frac{n}{nd_j})$, where tf_{ij} indicates the frequency of term j in the text i and nd_j denotes the total number of texts containing term j.

Given any two unit vectors **x** and **y** in R^m, the cosine similarity of **x** and **y** is defined by the inner product $\mathbf{x}^T\mathbf{y}$, i.e.,

$$\mathbf{x}^T\mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\|cos(\theta(\mathbf{x},\mathbf{y})) = cos(\theta(\mathbf{x},\mathbf{y})), \tag{1}$$

where $0 \leq \theta(\mathbf{x}, \mathbf{y}) \leq \pi/2$ denotes the angle of vectors **x** and **y**.

Let $\Pi = \{\pi_1, \pi_2, \dots, \pi_k\}$ denote a partitioning of the text vectors into k disjoint clusters such that

$$\bigcup_{k'=1}^{k} \pi_{k'} = \{\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_n}\}, \tag{2}$$

$$\pi_r \bigcap \pi_s = \phi, \forall r, s \in \Pi, r \neq s. \tag{3}$$

For each $k' \in 1, 2, \dots, k$, the centroid of the text vectors contained in the cluster $\pi_{k'}$ is

$$\mathbf{m}_{k'} = \frac{1}{n_{k'}} \sum_{\mathbf{x} \in \pi_{k'}} \mathbf{x}, \tag{4}$$

where $n_{k'}$ is the number of text vectors in $\pi_{k'}$. Then, the concept vector of the cluster $\pi_{k'}$ is defined as

$$\mathbf{c}_{k'} = \frac{\mathbf{m}_{k'}}{\|\mathbf{m}_{k'}\|}. \tag{5}$$

According to Cauchy–Schwarz inequality

Please cite this article as: C. Jia et al., Concept decompositions for short text clustering by identifying word communities, Pattern Recognition (2017), https://doi.org/10.1016/j.patcog.2017.09.045