Neurocomputing 158 (2015) 32–41

Contents lists available at ScienceDirect

Neurocomputing

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

# Chinese–Tibetan bilingual clustering based on random walk

Cheng-Xu Ye [a,b], Wu-Shao Wen [c,d,*], Chang-Dong Wang [e,d]

[a] School of Information Science and Technology, Sun Yat-Sen University, Guangzhou, PR China
[b] School of Computer Science, Qinghai Normal University, Xining, PR China
[c] School of Software, Sun Yat-Sen University, Guangzhou, PR China
[d] SYSU-CMU Shunde International Joint Research Institute (JRI), Shunde, PR China
[e] School of Mobile Information Engineering, Sun Yat-Sen University, Zhuhai, PR China

ARTICLE INFO

*Article history:*
Received 23 April 2014
Received in revised form
8 January 2015
Accepted 2 February 2015
Available online 18 February 2015

*Keywords:*
Multi-source clustering
Bilingual clustering
Chinese–Tibetan
Bilingual graph
Random walk

ABSTRACT

In recent years, multi-source clustering has received a significant amount of attention. Several multi-source clustering methods have been developed from different perspectives. In this paper, aiming at addressing the problem of Chinese–Tibetan bilingual document clustering, a novel bilingual clustering scheme is proposed, which can well capture both the intralingua document structures and interlingua document relations. The proposed scheme consists of three major phases. Firstly, to properly combine the feature structures of documents in different languages, a bilingual graph is constructed. In the second phase, two bilingual similarity matrices are computed based on the random walk performed in the bilingual graph. Finally, the similarity based clustering methods are performed on the two bilingual similarity matrices so as to generate cluster structures for documents in each language respectively, which lead to the corresponding bilingual clustering methods. Extensive experiments conducted on two Chinese–Tibetan bilingual document sets have confirmed the effectiveness of the proposed methods.



## 1. Introduction

In recent years, due to the development of the technology for data collection and storage, vast amount of data consisting of multiple sources have emerged. Such multi-source phenomenon is also called multi-view in some literatures [1–4]. In multi-source data, one common assumption is adopted that the class (cluster) structures in different sources would agree with each other, although they are characterized in different forms or different feature spaces. This implies that the clustering results in different sources should be compatible across different sources from the perspective of data clustering. Although it might be sufficient to learn the cluster structure from individual sources, when properly combining the multiple sources, they will complement each other and therefore improve the clustering performance. Such a task for finding compatible clustering results by combining different sources is termed multi-source clustering, which has received a significant amount of attention in the research field of clustering [2,3,5–8].

Several multi-source clustering methods have been proposed for addressing some specific multi-source clustering tasks [9–15]. For instance, in [12], to deal with the incomplete mapping between views, Eaton et al. proposed a multi-view algorithm based on constrained clustering, which propagates the pairwise constraints using a local similarity measure. To handle the case where each view would contain noise, Xia et al. [13] proposed a Markov chain method, based on which a robust multi-view spectral clustering is designed, which has a flavor of low-rank and sparse decomposition. For the case where some views may suffer from the missing of some data (i.e., examples withth some views missing), a partial multi-view clustering method was developed in [14]. The basic idea is to establish a latent subspace, where the instances w.r.t. the same example in different views are close to each other, and similar instances belonging to different examples in the same view should be well grouped. Similarly, a convex subspace representation learning approach was proposed for general multi-view clustering [15]. The basic idea is to identify a common subspace representation of the data across all views, based on which some standard clustering methods can be applied. However, due to the diversity of multi-source data, there is no common multi-source clustering method that is suitable for all multi-source clustering tasks.

For the Chinese–Tibetan documents, there exist some major diversities between two languages [16]. The first difference is the word order. For instance, the order of subject, predicate and object in Chinese is "subject+predicate+object", while the order of the three main sentence components in Tibetan is "subject+object+predicate". The second major difference lies in the morphological change, i.e., there is no morphological change in Chinese, while it is common to use morphological change in Tibetan. Another diversity is the means of sentence expression. In Chinese, the word order is the basis for

* Corresponding author at: SYSU-CMU Shunde International Joint Research Institute (JRI), Shunde, PR China. Tel.: +86 13924183355.
*E-mail addresses:* ycx@qhnu.edu.cn (C.-X. Ye), wenwsh@mail.sysu.edu.cn (W.-S. Wen), changdongwang@hotmail.com (C.-D. Wang).

http://dx.doi.org/10.1016/j.neucom.2015.02.003
0925-2312/

sentence expression, while the sentence expression in Tibetan mainly relies on the case-auxiliary word. The diversities between Chinese and Tibetan imply that the existing multilingual document clustering methods may be unsuitable for the case of Chinese–Tibetan bilingual document clustering. In this paper, aiming at addressing the task of clustering Chinese–Tibetan bilingual documents, we devise a novel bilingual graph, based on which some bilingual clustering methods are proposed.

In the proposed methods, to properly combine the feature structure of documents in different languages, we first construct a bilingual graph, which is a bipartite graph consisting of two parts, each representing one language. Apart from that, three types of edges are devised to respectively characterize the intralingua document similarities and interlingua document consistency. Based on the bilingual graph, two bilingual similarity matrices, one for each language, are computed based on the random walk performed in the bilingual graph. Compared with the original similarity matrix computed from the documents in single-language separately, the bilingual similarity matrix has an advantage that it can well capture the interlingua relation. Based on the bilingual similarity matrices, three representative similarity based clustering methods are performed, namely, *k*-centers [17], affinity propagation (AP) [18] and multi-exemplar affinity propagation (MEAP) [19], which lead to three bilingual clustering (bi-source clustering) methods termed bilingual *k*-centers, bilingual AP and bilingual MEAP respectively. Extensive experiments have been conducted to compare the proposed bilingual clustering methods with the single-source clustering methods as well as the existing multi-source clustering methods. Comparison results have confirmed the effectiveness of the proposed methods.

The remainder of the paper is organized as follows. Section 2 introduces the related work on multi-source clustering. The proposed bilingual clustering methods are described in detail in Section 3. Experimental results are reported in Section 4. We conclude our paper in Section 5.

## 2. Related work

In the literature of multi-source clustering, one of the earliest researches was conducted by Leih et al. [1]. In that work, to incorporate disparate forms of information from different sources, a Bayesian network approach is utilize to compute similarity matrix. Bickel and Scheffer [2] developed two types of multi-source clustering algorithms, namely, partitional multi-source clustering and hierarchical multi-source clustering, which are extended from the partitional clustering and hierarchical clustering respectively. Rogers et al. [5] developed a hierarchical non-parametric clustering model for multi-source data, where the lower level models each source separately with a flexible non-parametric mixture and the top level combines these mixtures to describe commonalities of the sources.

For the task of multilingual document clustering, Kim et al. [9] proposed a multi-source clustering that utilizes clustering results obtained in each source as a voting pattern in order to construct a new set of multi-source clusters. In [10], Taralova et al. developed a source constrained clustering (SCC) that extends *k*-means by enforcing each cluster to contain samples from a minimum number of sources. The SCC approach iterates between clustering and metric learning so as to minimize an objective function that takes into account grouping samples which are similar and representative of the sources. This SCC approach has been shown to generate good results in computer vision.

Matrix factorization is also utilized in devising multi-view clustering methods [11,20,21]. In [11], the problem of clustering vertices based on multiple graphs was addressed. The multiple graphs represent relations from different sources. A link matrix factorization is devised to combine information from multiple graph sources. In [20], Liu et al. proposed a NMF-based multi-view clustering algorithm by searching for the factorization that generates compatible clustering solutions across multiple views. In [21], Co-regularized Non-negative Matrix Factorization was developed, which extends NMF for multi-view clustering by jointly factorizing the multiple matrices through co-regularization.

Another widely investigated type of multi-source clustering is multi-source spectral clustering, due to the well-defined mathematical framework [3,4,22–24]. For instance, a multi-view spectral clustering was developed by generalizing the single-view normalized cut to the multi-view case, where the goal is to find a cut that is good enough on average while it may not be the best for a single graph [22]. This is a single-objective extension of the traditional single-source spectral clustering. On the other hand, a multi-objective formulation was developed in [24], which simultaneously considers the quality of a single cut on multiple graphs. In [3] and [4], another two spectral clustering frameworks have been developed for multi-view clustering, which respectively use co-training and co-regularization to make the clusterings agree with each other across different views.

## 3. The proposed methods

In this section, we describe in detail the proposed Chinese–Tibetan bilingual clustering (bi-source clustering) methods, which consist of three major phases. The first phase is to construct a bilingual graph such that the intralingua document structures and interlingua document relations can be well captured. Then, by performing random walks in the bilingual graph, two bilingual similarity matrices can be computed, one for documents in each language. Finally, similarity based clustering is performed on the bilingual similarity matrices to discover cluster structures for Chinese documents and Tibetan documents respectively.

### 3.1. Bilingual graph construction

Suppose that we are given a Chinese–Tibetan bilingual document set consisting of both Chinese documents $\mathcal{C}$ and the corresponding (translation) Tibetan documents $\mathcal{T}$, where $\mathcal{C}=\{\mathbf{x}_i \in R^d | i=1,\ldots,N\}$ and $\mathcal{T}=\{\mathbf{y}_i \in R^{d'} | i=1,\ldots,N\}$. Each pair of Chinese document $x_i$ and Tibetan document $y_i$ represents the same meaning, i.e., they are the same document in different languages. These two types of documents are represented as a "bag of words" using a TFIDF-based weighting scheme. Given such a bilingual document set, we need to construct some graph structure to properly combine the feature structures in different languages. To this end, a bilingual graph is devised, which is a bipartite graph consisting of two parts, each representing one language. Apart from that, there exist intralingua document similarity relation and interlingua document consistency, which form three types of edges of the bilingual graph.

First of all, we need to compute the similarity matrices of the documents in Chinese and Tibetan respectively. Let $\mathrm{sim}_C(\mathbf{x}_i,\mathbf{x}_j)$ denote the similarity between Chinese documents $x_i$ and $x_j$, which can be defined as the cosine similarity between the corresponding document feature vectors

$$\mathrm{sim}_C(\mathbf{x}_i,\mathbf{x}_j)=\frac{\mathbf{x}_i^\top \cdot \mathbf{x}_j}{\sqrt{\mathbf{x}_i^\top \cdot \mathbf{x}_i}\cdot\sqrt{\mathbf{x}_j^\top \cdot \mathbf{x}_j}} \tag{1}$$

It should be noticed that the proposed bilingual graph is not specific to the use of a particular similarity function like cosine similarity, but can also be used in conjunction with various similarity functions. The similarity matrix of the Chinese documents can be defined as $S^C=[\mathrm{sim}_C(\mathbf{x}_i,\mathbf{x}_j)]_{N\times N}$. Similarly, we can define and compute the similarity matrix of the Tibetan documents $S^T=[\mathrm{sim}_T(\mathbf{y}_i,\mathbf{y}_j)]_{N\times N}$. These two similarity matrices are used to characterize the document relation in