Information Processing and Management 44 (2008) 1397–1409

Contents lists available at ScienceDirect

Information Processing and Management

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

Towards effective document clustering: A constrained *K*-means based approach

Guobiao Hu [a,b], Shuigeng Zhou [a,b,*], Jihong Guan [c], Xiaohua Hu [d]

[a] *Department of Computer Science and Engineering, Fudan University, Shanghai 200433, China*
[b] *Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China*
[c] *Department of Computer Science and Technology, Tongji University, Shanghai 200092, China*
[d] *College of Information Science and Technology, Drexel University, Philadelphia, PA 19104, USA*

ARTICLE INFO

Article history:
Received 18 August 2007
Received in revised form 5 February 2008
Accepted 11 March 2008
Available online 25 April 2008

Keywords:
Document clustering
Semi-supervised learning
Spectral relaxation
Clustering with prior knowledge

ABSTRACT

Document clustering is an important tool for document collection organization and browsing. In real applications, some limited knowledge about cluster membership of a small number of documents is often available, such as some pairs of documents belonging to the same cluster. This kind of prior knowledge can be served as constraints for the clustering process. We integrate the constraints into the trace formulation of the sum of square Euclidean distance function of *K-means*. Then,the combined criterion function is transformed into trace maximization, which is further optimized by eigen-decomposition. Our experimental evaluation shows that the proposed semi-supervised clustering method can achieve better performance, compared to three existing methods.

1. Introduction

Document clustering is to partition a collection of documents into several clusters, such that the documents in the same cluster are as similar as possible, whereas those in any two different clusters are as dissimilar as possible. It has been applied to many fields such as data mining (Han & Kamber, 2001), information retrieval (Frakes & Baeza-Yates, 1992), topic detection and tracking (Allan, 2002).

Usually, document clustering is performed in unsupervised fashion, i.e, only unlabeled documents are handled. However, in real application scenarios, it is often the case that the users have some background knowledge about the dataset, which could be useful in the process of clustering (Wagstaff, Rogers, & Schroedl, 2001). This leads to a promising research direction, *semi-supervised clustering*.

Recently, semi-supervised clustering has attracted significant research effort in machine learning and data mining communities (Basu, Banerjee, & Mooney, 2002, 2004; Ji, Xu, & Zhu, 2006; Klein, Kamvar, & Manning, 2002; Wagstaff & Cardie, 2000; Wagstaff et al., 2001; Xing, Ng, Jordan, & Russell, 2002). Semi-supervised clustering uses additional constraints to guide the clustering process. The commonly used constraints are the pairwise constraints, which specify that two data items must be put into the same cluster (termed as *must-link*, simply *ML*) or two different clusters (referred to as *cannot-link*, simply *CL*). The exploitation of this kind of constraints can be dated back to Wagstaff and Cardie (2000), in which the authors incorporated *ML* and *CL* constraints into the COBWEB algorithm (Fisher, 1987) and achieved performance improvement.

The pairwise constraints under semi-supervised clustering can be used in two ways: *hard*, i.e., the constraints cannot be violated during the clustering process (Wagstaff et al., 2001, 2000), and *soft*, i.e., the constraints are used to help optimizing a

* Corresponding author. Address: Department of Computer Science and Engineering, Fudan University, Shanghai 200433, China.
E-mail addresses: gbhu@fudan.edu.cn (G. Hu), sgzhou@fudan.edu.cn (S. Zhou), jhguan@mail.tongji.edu.cn (J. Guan), thu@cis.drexel.edu (X. Hu).

0306-4573/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2008.03.001

criterion objective function (Basu, Bilenko, & Mooney, 2004; Ji et al., 2006). Our work in this paper falls into the second category.

Inspired by the work of Ji et al. (2006), in which *ML* constraints were integrated with Normalized Cut criterion function (Shi & Malik, 2000) and better performance was achieved, and based on the result of Zha, Ding, Gu, He, and Simon (2002), where the authors showed that the objective function of traditional *K-means* could be formulated as the trace expression of the original data matrix and then could be optimized further by eigen-decomposition, we incorporate the *ML* constraints into the trace expression of the *K-means* criterion function. Here, the constraints served as a penalty term to guide the optimization procedure are adapted to the semi-supervised clustering framework. The derived new approach is then applied to document clustering. In our implementation, we use the *vector space model* (VSM) (Salton & McGill, 1983) to represent the documents. In summary, our contributions in this paper are as follows:

(i) A novel semi-supervised clustering method, which incorporates the *ML* constraints into the trace expression of the *K-means* criterion function, is proposed and applied to document clustering. We term the proposed approach *S3–Kmeans*, which is an abbreviation of Semi-Supervised Spectral *K-means*.
(ii) Extensive experiments over seven different datasets are carried out, and the proposed approach is compared with three existing methods: the traditional *K-means*, the spectral relaxation based *K-means* (abbreviated as *S–Kmeans*) (Zha et al., 2002) and the semi-supervised clustering method *Cop–Kmeans*[1] (Wagstaff et al., 2001) that is also based on *K-means*. The experimental results show that our proposed method outperforms these three existing competitive methods.
(iii) As discussed in Davidson, Wagstaff, and Basu (2006), pairwise constraints are not always helpful for clustering, which is also demonstrated by *Cop–Kmeans* in this paper. However, the experimental results in this paper reveal that our proposed method achieves monotonically incremental performance improvement as the number of constraints increase, which implies that our method can make use of constraints more effectively than the existing methods, such as *Cop–Kmeans*.

The remainder of this paper is organized as follows. Section 2 surveys the related work. Section 3 gives the details of our proposed method. Section 4 presents the experimental settings and results. Section 5 concludes the paper and highlights future work.

2. Related work

Traditional document clustering is usually treated as an unsupervised learning task, i.e, only unlabeled documents are taken as input. Roughly, document clustering methods fall into three categories: partitioning, hierarchial, and graph-based (Zhao et al., 2001). Partitioning clustering is to directly split a dataset into k disjoint groups such that documents in the same group are more similar averagely with each other than those from any two different groups. *K-means* is the typical one of this kind. Hierarchial clustering proceeds successively by building a tree-like clusters structure, which can be done in either divisive (top-down) or agglomerative (down-up) way. Graph-based methods attempt to model documents as vertices of a weighted graph (Han, Karypis, Kumar, & Mobasher, 1998). The edge weight is determined by the similarity of the two corresponding documents. Then, the problem of document clustering is transformed to graph partitioning based on a certain criterion, such as Min Cut (Wu & Leahy, 1993), Ratio Cut (Chan, Schlag, & Zien, 1994), and Normalized Cut (Shi & Malik, 2000). Also, there exist some papers that model documents set and keywords set as two sets of component vertices in a bipartite graph (Zha, Ding, & Gu, 2001; Dhillon, 2001).

In practical scenarios, some prior knowledge is usually available, which could be helpful for clustering. How to use prior knowledge to help clustering is the core research issue of *semi-supervised clustering*. In the literature, extensive research on semi-supervised clustering has been done by the machine learning community, and the developed methods can be applied to document clustering, if the vector space model (*VSM*) is used for document representation. In what follows, we will focus on reviewing the related work of semi-supervised documents clustering.

Semi-supervised clustering is usually performed by imposing some constraints to an existing clustering method. These imposed constraints often come in the forms of *ML* and/or *CL* constraint pairs. As *K-means* algorithm (Hartigan & Wong, 1979) is a popular technique in data clustering for its simplicity and ease implementation/use, quite some research work has been done to take into account limited user supervision with *K-means* (Basu et al., 2002; Kulis, Basu, Dhillon, & Moony, 2005; Wagstaff et al., 2001;). Wagstaff et al. (2001) proposed a semi-supervised clustering method *Cop–Kmeans*, in which *ML* and *CL* constraints were incorporated into *K-means* and were not allowed to be violated during the clustering process. Basu et al. (2002) utilized a small number of labeled samples to generate initial centroids for *K-means*. A kernel-based method for semi-supervised *K-means* clustering was proposed in Kulis et al. (2005). It combines the sum of square Euclidean distance with the costs of violating *ML* and *CL* constraints. Though a powerful technique, the applicability of the kernel-based semi-supervised clustering is limited in practice, because the quality of clustering results is greatly affected by the chosen parameters (Yan & Domeniconi, 2006).

[1] We only consider the use of *ML* constraints in our experiments.