Data & Knowledge Engineering 68 (2009) 49–67

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

An active learning framework for semi-supervised document clustering with language modeling

Ruizhang Huang [a,*], Wai Lam [b]

[a] Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[b] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong

ARTICLE INFO

Article history:
Received 11 April 2007
Received in revised form 15 July 2008
Accepted 15 August 2008
Available online 16 September 2008

Keywords:
Document clustering
Semi-supervised
Active learning
Language modeling

ABSTRACT

This paper investigates a framework that actively selects informative document pairs for obtaining user feedback for semi-supervised document clustering. A gain-directed document pair selection method that measures how much we can learn by revealing judgments of selected document pairs is designed. We use the estimation of term co-occurrence probabilities as a clue for finding informative document pairs. Term co-occurrence probabilities are considered in the semi-supervised document clustering process to capture term-to-term dependence relationships. In the semi-supervised document clustering, each cluster is represented by a language model. We have conducted extensive experiments on several real-world corpora. The results demonstrate that our proposed framework is effective.

1. Introduction

Document clustering, dealing with the problem of grouping unlabeled text documents into a set of clusters, is of considerable interest for many applications. Sometimes, users may have their own preference for certain kinds of underlying clusters. For example, in the news document clustering task, a user can choose to group news documents according to topics of news events, such as "The 2008 Olympic Game", "South Korea Hostage", and "Nobel Prizes Awarded". Alternatively, another user can also choose to group news documents according to general categories, such as "sports", "finance", "technology", etc. Traditional document clustering methods cannot detect the user preference automatically. However, users can provide some information in the form of constraints to guide the clustering process. In this paper, we consider pairwise constraints as the type of user-provided information. Pairwise constraints contain must-link or cannot-link information between two documents indicating that two documents must be in the same cluster or must be in different clusters, respectively. Semi-supervised document clustering, which groups unlabeled documents into clusters taking into account the user-provided information, is a different problem from ordinary document clustering and has received much attention recently.

Existing semi-supervised document clustering approaches tackle the problem from the perspective on how to use the user-provided information. The user-provided information is generally useful for improving the performance of document clustering. Another factor that is able to help improve the document clustering performance is the quality of user-provided information. Most existing semi-supervised document clustering approaches involve user-provided information in a passive manner. Therefore, it becomes crucial for a user to provide the most "valuable" information. However, it is not feasible for a user to browse all text documents and select the most informative ones. A solution to this problem is to let the clustering approach play an active role in the process. Informative documents can be actively selected rather than chosen at random.

* Corresponding author.
E-mail address: rzhuang@se.cuhk.edu.hk (R. Huang).

0169-023X/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2008.08.008

In this paper, we aim to build an effective semi-supervised document clustering approach that automatically discovers informative constraints and finds a good document partition as well.

The first contribution of our approach is to automatically select informative document pairs for obtaining user judgments so that the clustering performance can be improved with as few supervised data as possible. Active learning, extensively studied in machine learning and applied to text classification, studies the closed-loop phenomenon of a learner selecting actions or making queries that influence what data are added to its training set [9]. We investigate an active learning approach so that the most informative document pairs can be chosen to form pairwise constraints. The active learning approach and semi-supervised document clustering are processed in an iterative manner. In the semi-supervised document clustering process, an optimal solution of document partition is discovered. However, this document partition is converged to a local optimal solution rather than a global one. Therefore, the active learning approach is incorporated to select document pairs according to the current intermediate clustering result so that the semi-supervised document clustering can find a better local optimal solution, hopefully the global optimal solution, in the next round. We investigate a gain function for selecting document pairs in our active learning process. This gain function is designed to measure how much information we can learn by revealing judgments of document pairs.

There has been little work on investigating active learning for semi-supervised document clustering. Basu et al. [3] proposed a two-step active learning scheme based on the farthest-first traversal. Pairwise constraints are discovered for finding good initialization of cluster centroids. In this method, active learning is a part of the preprocessing phase of the clustering. The active learning process and the semi-supervised document clustering process are not iteratively conducted. Pairwise constraints are discovered before the semi-supervised document clustering process and are not affected by the nature of the clustering result. Klein et al. [13] also considered active learning in semi-supervised clustering. However, this method does not deal with text document data. Active learning applied to semi-supervised document clustering is closely related to active learning applied to text classification. Nigam and McCallum [17] proposed an approach that employs active learning and EM to use unlabeled pool for training text classifiers. However, due to the nature of classification problem, an initial set of labeled documents or training documents are needed for *every* classifier. In our approach, we discover a small number of pairwise constraints for some, but not necessarily, all clusters.

The second contribution of our approach is that we investigate language modeling for representing clusters. Currently, most existing semi-supervised document clustering approaches are model-based clustering and can be treated as parametric models. Each cluster is represented by a set of parameters such as cluster centroids. For example, Constrained *K*-Means clustering algorithm [2], a recent semi-supervised document clustering approach, is derived from the *K*-Means document clustering algorithm. Cluster centroids are adopted for the cluster representation. Documents are assumed to follow multinomial distributions. These parametric models may not work well for the semi-supervised document clustering problem. The main reason is that the underlying clusters may not follow the distribution assumption of the parametric model. Language modeling, first introduced by Ponte and Croft [18], has been proposed as an alternative way to traditional models and has shown considerable success in information retrieval. The basic idea behind language modeling is formulated from solid statistical foundations and it often achieves promising results. One key idea of language modeling is to use a non-parametric probability distribution as the document representation.

Another contribution of our approach is to relax the "bag-of-words" assumption. In most of the semi-supervised clustering approaches, each text document is represented by a set of terms where a term refers to a single word or token. Terms in the documents are assumed to be independent. However, this "bag-of-words" assumption may not hold in reality. Different terms can be used to capture similar ideas due to different sources and different writing styles of authors. For example, "country" and "nation" are not independent and can sometimes be regarded as similar to each other. Therefore, it is advantageous if we can relax this assumption. Term-to-term dependence relationships can be captured by the term co-occurrence statistics. User-provided pairwise constraints contain useful information for estimating the term-to-term dependence relationships especially for the terms related to the same cluster. Terms co-occurred frequently in the documents labeled with the same cluster should have a higher probability of cohesiveness. Terms co-occurred frequently in the documents labeled with different clusters are relatively not so discriminative. We design the gain function for actively selecting document pairs to be helpful for estimating term co-occurrence probabilities. Pairwise constraints are then updated by the selected document pairs and are used to guide the subsequent semi-supervised document clustering process.

We have conducted extensive experiments on our proposed approach to evaluate: (1) whether our active learning approach is effective; (2) whether representing clusters with language modeling is useful. Therefore, the semi-supervised document clustering approach proposed by Basu et al. [3] is investigated for comparison as it is a recent method on active learning in semi-supervised document clustering. We also compared our approach with a model-based semi-supervised document clustering approach [3] which uses centroids to represent clusters and does not learn constraints. Experimental results demonstrate that our framework is more effective.

The remaining parts of this paper are organized as follows. Section 2 reviews related work on document clustering and semi-supervised document clustering. In Section 3, we present our semi-supervised document clustering approach which uses the discovered pairwise constraints to aid the document partition. In Section 4, we describe our proposed active learning approach for discovering informative document pairs. In Section 5, we briefly discuss the overall system design of our framework. Empirical results are presented in Section 6. We finally present conclusions and future work in Section 7.