Pattern Recognition Letters 52 (2015) 25–31

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Interactive textual feature selection for consensus clustering ☆

Geraldo N. Corrêa [a], Ricardo M. Marcacini [b,*], Eduardo R. Hruschka [a], Solange O. Rezende [a]

[a] Institute of Mathematical and Computer Science, University of São Paulo, São Carlos, SP, Brazil
[b] Federal University of Mato Grosso do Sul, CPTL, Três Lagoas, MS, Brazil

ARTICLE INFO

Article history:
Received 13 December 2013
Available online 21 October 2014

Keywords:
Interactive feature selection
Consensus clustering
Text mining

ABSTRACT

Consensus clustering and interactive feature selection are very useful methods to extract and manage knowledge from texts. While consensus clustering allows the aggregation of different clustering solutions into a single robust clustering solution, the interactive feature selection facilitates the incorporation of the users' experience in the clustering tasks by selecting a set of textual features, *i.e.*, including user's supervision at the term-level. We propose an approach for incorporating interactive textual feature selection into consensus clustering. Experimental results on several text collections demonstrate that our approach significantly improves consensus clustering accuracy, even when only few textual features are selected by the users.

1. Introduction

Text clustering methods are very useful for automatic organization of documents into clusters, where more similar documents are found in the same cluster and are separated from more dissimilar documents [1]. This organization provides intuitive browsing of large text collections and facilitates the exploratory analysis of unknown data to reveal implicit knowledge from texts [2].

There is a wide variety of clustering methods, such as partitional and hierarchical clustering [3], density-based clustering [4], graph-based clustering [5], spectral clustering [6], co-clustering [7], model-based clustering [8], and fuzzy clustering [9]. Furthermore, each clustering method has its own distinct biases that influence how the clusters are identified within the textual data [10]. Despite this variety, no specific clustering algorithm is able to identify all the shapes and clustering structures [11–13,3]. In this sense, consensus clustering allows the combination of different clustering solutions into a unique and more robust clustering solution [10,13,14]. Thus, if a document is mistakenly allocated to a particular clustering solution, the same document will not necessarily be mistakenly allocated in other clustering solutions — *i.e.*, eventual errors can be corrected in the final solution obtained by consensus clustering [15].

Although consensus clustering can provide more robust data partitions, the unsupervised organization of textual collections presents some drawbacks about the understanding of the generated clusters [16]. In general, the data partitions are obtained based only on distance measures between documents, which often do not capture the notion of proximity expected by users [16]. Semi-supervised clustering algorithms attempt to mitigate this problem by using a set of constraints to indicate which documents should be in the same cluster (must-link constraint) or in different clusters (cannot-link constraint) [17,18]. However, providing a reasonable set of constraints is a very difficult task for the users, since usually there is no prior knowledge about the spatial structure of the data [19].

The interactive feature selection is a promising way to include the user's experience in text clustering tasks [16,20,21]. Unlike approaches that require a set of constraints, the interactive feature selection presents an initial clustering solution to the user, where each cluster has a set of associated textual features (words, phrases or expressions). Users can then indicate the textual features that are more interesting according to their experience and intuitions about the problem domain — knowledge of spatial structure of the data can still be used, but it is not required. After users' interaction, the clusters are refined considering the selected textual features, thereby favoring a clustering solution closer to the user's expectations. Note that active learning techniques can be used during the interactive feature selection process to minimize the number of queries for the user and to provide a suitable set of textual features for each cluster [22].

In this paper, we introduce a novel approach for consensus clustering with interactive feature selection (CCIFS for short). A preliminary study was presented in Ref. [23]. While existing approaches require the use and adaptation of a particular clustering algorithm to incorporate interactive feature selection, CCIFS allows the use of interactive feature selection in any text clustering algorithm, thereby enabling the inclusion of users' feedback in a robust strategy of consensus clustering.

The key idea of CCIFS is to represent the text collection by two data views: (1) low-level features and (2) high-level features. The low-level

☆ This paper has been recommended for acceptance by Y. Chang.

* Corresponding author. Tel.: +55 (16)3373 9659; fax: +55 (16)3373 9751.
E-mail address: ricardo.marcacini@ufms.br (R.M. Marcacini).

http://dx.doi.org/10.1016/j.patrec.2014.09.008
0167-8655/

features consist of the traditional bag-of-words model, which simply associates single words and their frequencies to represent text documents. The interactive feature selection is applied to extract, according to the user's experience, the high-level features. Unlike existing approaches [16,19,21], where the users' feedback is used only to redefine the weight of single words in the bag-of-words, CCIFS identifies correlated words to compose high-level features. For example, if "artificial", "neural" and "network" are three words existing in the bag-of-words, then the interactive feature selection can extract the new (high-level) feature "artificial neural network". The set of correlated words selected by the users' feedback from the interactive feature selection is used to compose the high-level features data view, which complements the traditional bag-of-words model. After the extraction of the two data views, several clustering solutions are obtained for each data view and, finally, the clusters are combined into a single clustering solution using consensus clustering.

The main contribution of our work is the exploration of how the interactive feature selection can be incorporated effectively into robust text clustering tasks. CCIFS achieves this by defining the contribution factor of each data view during the consensus clustering. A thorough experimental evaluation, using nine real-world textual collections, was carried out to analyze the improvements obtained when interactive feature selection is incorporated into consensus clustering. We compare three different scenarios: (i) consensus clustering without interactive feature selection, *i.e.*, using only low-level features, (ii) consensus clustering using both data views (low-level features and high-level features), and (iii) consensus clustering using only high-level features. The first scenario represents a traditional approach for consensus clustering. The second and third scenarios represent the use of interactive feature selection for consensus clustering introduced in this paper. Statistical analysis of the experimental results reveal that the CCIFS obtains better clustering solutions when both data views are used in the consensus clustering (scenario ii), even when only a few textual features are selected by the users.

The remainder of this paper is organized as follows. The next section presents related work on the use of interactive feature selection for text clustering tasks. Section 3 describes the proposed method for consensus clustering with interactive feature selection. An experimental evaluation is carried out and the results are discussed in Section 4. Finally, Section 5 presents conclusions and future works.

2. Related work

Interactive feature selection for textual data was first introduced by Raghavan et al. [22], in which an active learning technique is used to identify a set of potential relevant textual features. The users provide feedback about the most important features according to their experience on the problem domain. The authors used a text classification task to assess the effectiveness of interactive feature selection. The experimental results showed that humans have good intuition to identify relevant textual features, even when users are not domain experts.

Recently, interactive feature selection has been proposed for text clustering tasks. In Ref. [16], the authors describe a new approach to improve the clustering accuracy by including human supervision at the textual feature level. First, the approach presents the highly ranked features to the users, *i.e.*, features with the highest weight (*e.g.* frequency values) in each cluster from an initial clustering. Then, the users have to label each textual feature as either "accept" or "don't know" — according to their understanding about the textual collection. The *accepted* features and a number of highly ranked features are used to compose a new document representation. Finally, a clustering algorithm iterates using the new document representation, thereby producing clusters that hopefully match the user's expectations.

Since the quality and interpretation of textual features are significant issues for interactive feature selection (from the user's perspective), a text clustering approach with human supervision at term-level called AL^{2}FIC (active learning to frequent itemset-based clustering) was proposed by Marcacini et al. [20]. In this approach, the textual features presented to the users are formed by a set of correlated words that are more interpretable than single words of the bag-of-words model. The correlated words are extracted with the use of algorithms for finding frequent itemsets — *e.g.*, with the well-known Apriori algorithm [24]. Several studies indicate that text clustering tasks based on frequent itemsets are more suitable for the interpretation of the clustering structure [25–27]. Thus, when the user selects a textual feature (frequent itemset), (s)he is also providing feedback on other similar features and related concepts. Moreover, the AL^{2}FIC uses an active learning technique to present only the most representative frequent itemsets, thereby minimizing the total number of users' queries needed to increase the clustering accuracy.

A common characteristic of the approaches presented above is the use of interactive feature selection to refine the document representation. The basic idea is that if the users can influence the document representation, according to their knowledge about the problem domain, then the generated clustering solution is closer to the users' expectations. However, the clustering solution is still dependent on a particular clustering algorithm, which may present a bias that does not satisfy the users' interest. In addition, the information contained in the initial documents representation (original features) is overlooked during later iterations of the clustering algorithm. Thus, a bad initial selection of textual features can even lead to a cluster solution with low accuracy [20].

As discussed in Ref. [3], the use of consensus clustering is a promising way to alleviate these drawbacks. First, it is well known that the combination of different clustering solutions can yield to a more robust clustering solution. Second, a new document representation, extracted by interactive feature selection, can be used as an alternative textual data view, thereby complementing the initial feature set. Some experimental studies on clustering with multi-view data show that combining clusters from two or more views of the same dataset can lead to superior data partitions [28,29]. These observations have motivated us to employ related approaches to incorporate interactive feature selection into a robust consensus clustering framework. To the best of our knowledge, this aspect has not been addressed in the literature.

3. Consensus clustering with interactive feature selection

Our approach (CCIFS) can be divided into two main steps: (1) high-level features extraction and (2) clustering from high-level and low-level features. The first step uses interactive feature selection to identify a relevant set of features for the clustering task — according to the user's feedback. The selected features are used to form an alternative data view of the document collection (called high-level features), where each document has a relevance value associated to each high-level feature. The original feature set from the bag-of-words model is maintained during the clustering process as a data view that is called "low-level features". In the second step, multiple data partitions are obtained from both data views (high-level and low-level features). The data partitions are combined into a single and hopefully more robust clustering solution, in which it is possible to define the contribution factor of each data view for the consensus clustering.

3.1. High-level features extraction

We adopt the AL^{2}FIC [20] to perform the interactive feature selection. Thus, the users provide feedback on the frequent itemsets that best represent the textual collection. A frequent itemset is a set of words that co-occur in documents morethan a given threshold