Improving Consensus Clustering of Texts Using Interactive Feature Selection[*]

Ricardo M. Marcacini
Mathematical and Computer Sciences Institute
University of São Paulo - Brazil
rmm@icmc.usp.br

Marcos A. Domingues
Mathematical and Computer Sciences Institute
University of São Paulo - Brazil
mad@icmc.usp.br

Solange O. Rezende
Mathematical and Computer Sciences Institute
University of São Paulo - Brazil
solange@icmc.usp.br

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 text clustering tasks by selecting a set of high-level features. In this paper, we propose an approach to improve the robustness of consensus clustering using interactive feature selection. We have reported some experimental results on real-world datasets that show the effectiveness of our approach.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*

Keywords

Consensus Clustering; Interactive Feature Selection

1. INTRODUCTION

Consensus clustering and interactive feature selection have emerged as useful methods to organize and manage the implicit knowledge from text collections. Consensus clustering combines different clustering solutions from a particular text collection into a single robust clustering solution [5]. Recent studies show that the use of consensus clustering presents promising results in various relevant problems for text clustering, such as the knowledge reuse and clustering of heterogeneous data [1]. Interactive feature selection uses active learning to insert external knowledge in text clustering tasks [2], where users provide feedback by selecting adequate features for each cluster. The selected features are used to refine the clustering solution according to users' expectations and to define meaningful cluster labels. A major advantage of interactive feature selection is that it requires less effort from users than other methods of active learning [4]. Moreover, it is more natural for users to provide feedback on textual features than to indicate a set of constraints which is required for semi-supervised clustering [3].

In this paper we present a study on improving the consensus clustering robustness using interactive feature selection. Unlike existing methods, where a particular clustering algorithm is adapted to incorporate interactive feature selection, we propose a general approach that allows the use of any distance-based clustering algorithm. In our approach, the interactive feature selection is applied to extract, according to the user experience, an additional view of textual data called high-level document-feature matrix. This additional data view is used to complement the view provided by the traditional bag-of-words model – the low-level document-feature matrix – which simply associates words and their frequencies to represent text documents. Several data partitions are obtained for each data view by using various runs of clustering algorithms and, finally, the clusters are combined into a single clustering solution using consensus clustering.

We carried out an experimental evaluation on real world text collections to demonstrate the effectiveness of the proposed approach. The results show that the incorporation of the high-level data view (obtained from interactive feature selection) in the consensus clustering significantly improves the overall clustering accuracy.

2. INTERACTIVE FEATURE SELECTION FOR CONSENSUS CLUSTERING

For convenience, consider that each document i of a given textual collection is described by a vector of m feature values $d_i = (t_1, t_2, ..., t_m)$, where each feature is a single word that occurs in texts and its value is based on the frequency of occurrence of the word in the document. The text collection $D = \{d_1, d_2, ..., d_n\}$ contains a set of documents and represents the bag-of-words model, namely the low-level document-feature matrix. While the semantic information about words are ignored in the bag-of-words model, the clustering quality can be significantly improved when considering higher-level information about the textual data. In this sense, the goal of our interactive feature selection is to extract high-level features composed of correlated words.

We explore the well known algorithms for mining frequent itemsets to identify correlated words. First, we extract a set of frequent itemsets using the Apriori algorithm from the text collection D. We represent each frequent itemset as a triple $f = (C, \vec{LS}, T)$, where C is the document set covered by the frequent itemset, $\vec{LS}$ is the linear weighted sum of the documents related to the itemset, i.e., $\sum_{d \in C} \frac{d}{|C|}$, and T is the set of the correlated words that identifies the frequent itemset. Next, the extracted frequent itemsets are

[*]The authors wish to thank FAPESP for financial support (process 2010/20564-8, 2011/19850-9, and 2012/13830-9)

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

summarized in k clusters by running our algorithm for frequent itemset-based clustering called AL^{2}FIC [3]. In this algorithm, an active learning procedure supports users to select the best itemsets of each cluster, in which the user can analyze if the correlated word set T of a frequent itemset is appropriate to label a particular cluster. Finally, the frequent itemsets selected by the user are used to obtain the high-level document-feature matrix. In our approach, the feature value $w_{(d_i,f_j)}$ of a frequent itemset f_i for the document d_i is calculated by the cosine similarity between the document and the linear weighted sum of the frequent itemset, i.e., $w_{(d_i,f_j)} = cos(d_i, \vec{LS})$.

At this stage of our approach, the text collection is represented by two views: the low-level document-feature matrix and the high-level document-feature matrix. For each data view, data partitions are obtained by running various clustering algorithms or alternatively repeated runs of the same algorithm with different parameter values. It is important to note that, unlike approaches that specialize a particular clustering algorithm to incorporate interactive feature selection, we can use any distance-based clustering algorithm.

In order to perform the consensus clustering, the generated data partitions are aggregated by means of a co-association matrix. The basic idea is to summarize the p data partitions from a particular data view in a matrix whose elements are $M(i,j) = \frac{a_{ij}}{p}$, where a_{ij} is the number of times that documents d_i and d_j are in the same cluster. Thus, let M_L the co-association matrix obtained from data partitions of low-level features, and let M_H the co-association matrix obtained from data partitions of high-level features. We compute the final co-association matrix M_F using the following equation:

$$M_F(i,j) = \alpha M_L(i,j) + (1-\alpha)M_H(i,j) \qquad (1)$$

for all documents d_i and d_j, where α $(0 \leq \alpha \leq 1)$ indicates the contribution factor (weight) of each view for the consensus clustering. The co-association matrix represents a new similarity matrix between the text documents. In order to finish this step, the UPGMA clustering algorithm [1] is used to obtain a final clustering solution from the co-association matrix M_F, which represents the consensus clustering using information extracted from interactive feature selection.

3. EXPERIMENTAL EVALUATION

We carried out an experimental evaluation using six text collections composed of computer science papers from ACM Digital Library. Each text collection contains approximately 500 articles organized into 5 clusters. We compared three strategies of consensus clustering:

1. Only low-level features: consensus clustering obtained using only the co-association matrix M_L, i.e., the contribution factor value is $\alpha = 0$.
2. Only high-level features: consensus clustering obtained using only the co-association matrix M_H, i.e., the contribution factor value is $\alpha = 1$.
3. Combining high and low-level features: consensus clustering from the two co-association matrices, i.e., the contribution factor value is in the range $0 < \alpha < 1$.

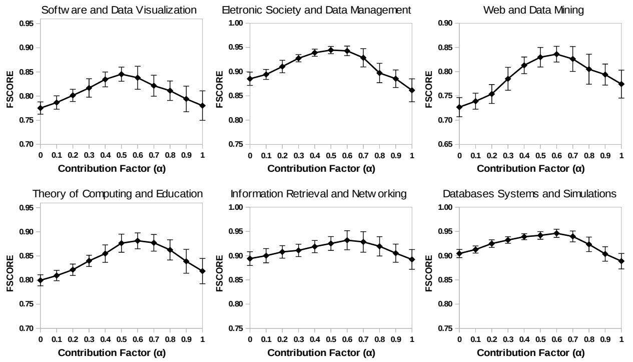

Figure 1: Consensus clustering accuracy for each text collection according to the contribution factor.

Several data partitions were generated using different runs of the k-means algorithm. For the interactive feature selection, we apply simulated users [2, 3] to select approximately 100 high-level features. Figure 1 illustrates the accuracy (F_{SCORE} index [1]) of the consensus clustering for each textual collection. Statistical analysis of these results (95% confidence level) reveals that the combination of high and low-level features achieved superior results, in which a significant improvement in accuracy is obtained when the contribution factor is close to 0.5 or 0.6. On the other hand, we did not observe significant improvement in the consensus clustering accuracy when the contribution factor of the two views is very unbalanced, i.e., for α close to 0 or 1.

4. CONCLUDING REMARKS

In this paper, we present an approach that explores interactive feature selection to improve the consensus clustering robustness. The features are selected using an active learning technique based on frequent itemsets to extract a high-level document-feature matrix, which complements the traditional bag-of-words model. The proposed approach is potentially useful for several real-world applications, especially when domain experts can participate in the knowledge extraction process by providing high-level features. Directions for future work involve the use of advanced text preprocessing techniques with interactive feature selection, such as named entity recognition and bag-of-concepts, for extracting alternative data views from texts.

More details on the experimental evaluation, as well as the text collections and algorithms used in this work, are available at http://sites.labic.icmc.usp.br/torch/www2013/.

5. REFERENCES

[1] D. C. Anastasiu, A. Tagarelli, and G. Karypis. Document Clustering: The Next Frontier. Technical Report. University of Minnesota, 2013.

[2] Y. Hu, E. E. Milios, and J. Blustein. Interactive feature selection for document clustering. In *ACM Symposium on Applied Computing*, pages 1143–1150, 2011.

[3] R. M. Marcacini, G. N. Correa, and S. O. Rezende. An active learning approach to frequent itemset-based text clustering. In *21st ICPR*, pages 3529 –3532, 2012.

[4] H. Raghavan, O. Madani, and R. Jones. InterActive feature selection. In *19th IJCAI*, pages 841–846, 2005.

[5] A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2003.