21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

An Active Learning Approach to Frequent Itemset-Based Text Clustering

Ricardo M. Marcacini, Geraldo N. Corrêa and Solange O. Rezende
Mathematical and Computer Sciences Institute - ICMC
University of São Paulo - USP - São Carlos, SP, Brazil
rmm@icmc.usp.br, geraldo@feb.br, solange@icmc.usp.br

Abstract

Frequent itemset-based text clustering has emerged as a promising way to automatic organization of text documents, because it allows high clustering accuracy combined with understandable cluster descriptors. However, the clustering results may not be satisfactory because they do not reflect the user's point of view. In this context, active learning is an interesting approach to incorporate the user's knowledge in the text clustering task by querying the users about the data. We introduce an active learning approach to frequent itemset-based text clustering called AL^{2}FIC. In our approach, the users can provide feedback directly on the cluster descriptors without the need to know the document labels. An experimental evaluation on real text collections demonstrated that our AL^{2}FIC approach significantly increases the text clustering performance even when only few descriptors are selected by the users.

1. Introduction

Frequent itemset-based text clustering, introduced in [5], has emerged as a promising way to automatic organization of text documents, because it allows high clustering accuracy combined with understandable cluster descriptors [1]. A frequent itemset is a set of words, which co-occur in documents more than a given threshold value called minimum support. The key idea behind frequent itemset-based text clustering is that documents belonging to the same cluster probably share more common itemsets than documents from different clusters [8]. Moreover, the frequent itemsets are used as cluster descriptors, providing intuitive interpretation and browsing of the clustering results [18].

Over the years, various text clustering algorithms based on frequent itemsets have been proposed in the literature [5, 8, 14, 18]. These algorithms allow an unsupervised process to obtain the clusters, therefore minimizing the users effort. However, the results may not be satisfactory because they do not reflect the user's point of view, as well as their experiences and expectations [12]. In this context, active learning is an interesting approach to incorporate the user's knowledge in the text clustering task by querying the users about the data [11]. The queries are done by requesting more information only on documents that are difficult to cluster or more pertinent to the clustering goal.

In clustering tasks, active learning has been used to support the semi-supervised clustering [4]. In this case, the user selects a set of "must-link" and "cannot-link" pairs of documents. This external information is used to reassign documents to the clustering process [3, 4, 12, 10], to learn about a more appropriate distance metric measure [6], and to select the most relevant features [16, 9]. Other studies use active learning to select a labeled document sample for a supervised learning task [13].

A serious disadvantage of these approaches is that users are required to provide a reasonable large set of "must-link" and "cannot-link" constraints, or a labeled document sample [9]. This is a difficult task for the users, since it is necessary to know the true labels of the documents. Furthermore, active learning is normally used in classification tasks, where a training set is available and the class labels are known. However, in the clustering context there is no previous knowledge about the cluster labels. In this sense, clustering using frequent itemsets is an interesting way to incorporate the user's experience in the clustering process. The users can provide feedback directly on the cluster descriptors without the need to know the document labels.

In this paper, we introduce an approach called AL^{2}FIC (Active Learning to Frequent Itemset-based Text Clustering). In our AL^{2}FIC, the best top-q frequent itemsets of each cluster (candidate descriptors) are presented to the users. Unlike existing approaches, a set of constraints or a labeled document sample is not required, but users can incorporate their knowledge by selecting some descriptors for the clusters. For this pur-

978-4-9906441-0-9 ©2012 ICPR

pose, we developed a clustering technique based on the EM algorithm [7], which iteratively updates the cluster descriptors according to the users selections. In each iteration the list of the best top-q frequent itemsets is updated and presented to the user. We carried out an experimental evaluation on real text collections and the results demonstrated that our AL^{2}FIC approach significantly increases the text clustering performance even when only few descriptors are selected by the users.

2. Active Learning to Frequent Itemset-Based Text Clustering (AL^{2}FIC)

We adopt the vector-space model [17] for structured representation of texts. In this model, given a text collection D with r documents, $D = \{d_1, d_2, ..., d_r\}$, each document d is represented by a vector of terms

$$d = \{t_1, t_2, ..., t_l\} \qquad (1)$$

where t_i is a value that indicates the importance of the term i in the document. A binary version of these vectors is used to obtain a set of frequent itemsets $F = \{f_1, f_2, ..., f_m\}$. We used the Apriori algorithm [2] to extract the frequent itemsets from the texts.

In our approach, each frequent itemset f also has a representation in the vector-space model. Considering that f is a frequent itemset and S_f the set of documents in which the itemset f occurs, then f_V represents the mean vector of the documents belonging to S_f (Equation 2).

$$f_V = \frac{1}{|S_f|} \sum_{\forall d \in S_f} d \qquad (2)$$

A cluster set A of frequent itemsets is similiarly represented by a vector c (cluster center) with the mean vectors of the frequent itemsets belonging to A, according to Equation 3.

$$c = \frac{1}{|A|} \sum_{\forall f_V \in A} f_V \qquad (3)$$

At this point, each object of the clustering task is defined in the same vector-space model. Thus, it is possible to compute the proximity between any two objects by using the cosine similarity (Equation 4), where x_i can be (i) documents d, (ii) frequent itemsets f_V, or (iii) cluster centers c.

$$cos(x_1, x_2) = \frac{x_1 \cdot x_2}{\parallel x_1 \parallel \parallel x_2 \parallel} \qquad (4)$$

The cosine similarity is easy to understand and is the most well known similarity function used in the text domains [1].

After defining the structured representation of texts and the similarity measure, we now define the clustering process. The clustering strategy of our AL^{2}IFC approach has two steps: (1) Initial Clustering and (2) Clustering Refinement. In the first step, a clustering algorithm based on EM (Expectation Maximization) such as k-means is applied to obtain an initial clustering $P = \{A_1, A_2, ..., A_k\}$ from the frequent itemsets $F = \{f_1, f_2, ..., f_m\}$. Each cluster $A_i \in P$ contains frequent itemsets such that $A_1 \cup A_2 \cup ... \cup A_k = F$ and $A_1 \cap A_2 \cap ... \cap A_k = \emptyset$. Then, the vector representations are obtained for each cluster by using the Equation 3.

Algorithm 1: AL^{2}FIC - Clustering Refinement

```
Input:
    $P$: initial clustering of frequent itemsets
    $q$: size of the top-q frequent itemsets
    $maxQueries$: maximum number of user's queries
Output:
    $P^{[new]}$: refined clustering
1 repeat
2     foreach cluster $c \in P$ do
3         compute a top-q list $L$ of the frequent itemsets in the
          cluster $c$;
4         user selects one frequent itemset $f \in L$;
5         perform a reweighting of the cluster center $c$ from
          selected itemset $f$;
6     end
7 until $maxQueries$ is reached;
8 return $P$;
```

In the second step of the clustering strategy, active learning is used to refine the initial clustering according to the user's experience, as shown in Algorithm 1. In this algorithm, users refine the clustering through their descriptors in an iterative way. For each cluster, a list of the best itemsets is presented to the user (line 3). The itemsets are ranked according to the coverage $|S_f|$ (number of documents in which the itemset occurs) and the proximity to the cluster center $cos(f_V, c)$, as described in Equation 5. Thus, if an itemset f has high coverage and belongs to the appropriate cluster, then f will get a high ranking score.

$$score(f_V, c) = |S_f| * cos(f_V, c) \qquad (5)$$

The user must select one itemset from the list, *i.e.*, choose the best descriptor based on their prior knowledge about the data (line 4). Considering that $f_V^{[sel]}$ is the mean vector of the itemset selected by the user (see Equation 2), then the cluster center is updated according to Equation 6, where the vector of the cluster center is changed to meet the user's expectations (line 5).

$$c^{[new]} = c + f_V^{[sel]} \qquad (6)$$

The motivation of the cluster center reweighting is as follows: with the addition operation between the vectors c and $f_V^{[sel]}$, the cluster center is moved in the $f_V^{[sel]}$ direction, *i.e.*, the new center $c^{[new]}$ is calculated to meet the user's expections represented by the $f_V^{[sel]}$. Thus, subsequent iterations of the EM are based on the new center. This process continues until the maximum number of queries is reached (line 7).

$$label(d) = \underset{i \in \{1..k\}}{\operatorname{argmax}} \; cos(d, c_i) \qquad (7)$$

Finally, after the clustering refinement by the user, each document is assigned to the nearest cluster center (Equation 7), by using the cosine similarity. Thus, each document of the text collection receives a cluster label, thereby obtaining a final text clustering solution.

3. Experiments and Results

In this section, we present and discuss an experimental evaluation of our AL^{2}IFC approach. For this purpose, we use six text collections from various sources. The main objective is to analyze the impact of active learning for frequent itemset-based text clustering. We compared the initial clustering obtained by unsupervised k-means algorithm with the clustering after the incorporation of knowledge by the users. Table 1 presents details of the text collections used in the experiment.

Text Collection	#Documents	#Terms	#Classes
ACM	394	1555	5
Hitech	2301	2289	6
LATimes	6279	6141	6
NSF	10521	10160	16
Re8	7674	7555	8
Reviews	4069	4047	5

Table 1. Summary of text collections used in the experimental evaluation.

The lowest text collection has 394 documents, while the largest has 10521 documents. All the texts were preprocessed following the recommendations in [15]: (i) stopwords removal, such as pronouns, articles and prepositions, and (ii) term stemming by Porter algorithm. For each text collection, we extracted frequent itemsets with 2-terms, by using the Apriori algorithm. The number of clusters k has been configured according to the number of classes of the text collections.

3.1 Simulating user interactions

In our experiments, we simulated the user's selections of frequent itemsets for each cluster. The simulation is based on the classes of the documents. Thus, when the AL^{2}FIC presents the top-q frequent itemsets for each cluster, we apply the InfoGain measure [17] to select the best itemset according to the classes of the text collections. These classes are considered to be the clusters expected by the user.

The simulation of the user's interactions is very important because we need to run AL^{2}FIC several times to perform a statistical comparison. This requires a great effort when performed under human supervision. In addition, the simulation of the users is a way to allow the future replication of the experiment.

3.2 Results

In this section, we present the clustering results obtained with our AL^{2}FIC approach. The initial clustering (step 1) was obtained with 30 different initializations of the k-means algorithm. In the clustering refinement (step 2), we presented the top 15 frequent itemsets to the user. The maximum number of queries were set equal to the number of clusters. Thus, it was possible to select only one frequent itemset for each cluster. Finally, the well-known FScore measure [17] was used to assess the clustering quality.

Figure 1 shows a comparison of the clustering quality with the AL^{2}FIC approach. The initial clustering represents a traditional approach, while the cluster refinement represents clustering improvement obtained with the use of active learning. We compared statistically several runs of our AL^{2}FIC. According to the nonparametric Wilcoxon matched-pairs signed-ranks test, the AL^{2}FIC approach significantly improves the clustering quality (with 95% confidence interval).

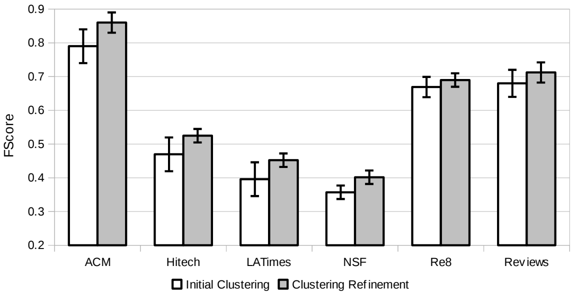

Figure 1. Clustering improvement with the use of active learning

It is important to note that the performance of the use of active learning is dependent on the characteristics of each text collection. Furthermore, we observe that the quality of clustering refinement is influenced by the quality of the initial clustering. A bad initial clustering provides a *top-q* list of unrepresentative frequent itemsets for the clusters. However, in most cases the

AL^{2}FIC approach is an intuitive way for users to incorporate their experiences. To illustrate, Figure 2 shows the best frequent itemsets for the class "Embedded Network" of the ACM text collection. The FScore values are presented in each step of the AL^{2}IFC: (a) initial clustering and (b) clustering refinement.

(a) Initial Clustering	**(b) Clustering Refinement**
{node,graph}	{node,graph}
{node,network}	{network,messag}
{algorithm,time}	{node,network}
{node,number}	{comput,graph}
{time,node}	{time,sensor}
{data,node}	{node,messag}
{time,set}	{network,packet}
{time,number}	{data,network}
Class FScore: 0.59	**Class FScore: 0.68**

Figure 2. Cluster descriptors obtained for the "embedded network" class of the ACM text collection

The AL^{2}FIC software and the text collections used in the experimental evaluation are available in the website `http://sites.labic.icmc.usp.br/torch/al2fic/`.

4 Concluding remarks

In this paper, we proposed an active learning approach to frequent itemset-based text clustering. In existing approaches, the user must provide a labeled document sample or a set of constraints. In our AL^{2}FIC, the user can provide a feedback through the cluster descriptors (frequent itemsets), which is more natural for the users. This is the main contribution of our approach. The experimental evaluation provides evidence that our approach improves the clustering quality, even with few queries for the users (only one query for each cluster). The results are interesting and the proposed approach is potentially useful for other pattern recognition tasks such as interactive feature selection and for bridging the semantic gap between low-level text features and the high-level human concepts.

Some well-known challenges for (i) determining the true number of clusters k and (ii) the minimal support for generation of frequent itemsets are also relevant to the performance of the AL^{2}FIC approach. However, the investigation of the influence of these parameters is out of the scope of this paper, and it will be addressed in future work. In addition, we also plan to evaluate the AL^{2}FIC approach using human supervision.

Acknowledgements. The authors wish to thank FAPESP for financial support (process 2010/20564-8 and 2011/19850-9).

References

[1] C. C. Aggarwal and C. Zhai. A survey of text clustering algorithms. In *Mining Text Data*, pages 77–128. Springer-Verlag, 2012.

[2] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *20th Int. Conf. on Very Large Data Bases (VLDB)*, volume 1215, pages 487–499, 1994.

[3] S. Basu, A. Banerjee, and R. Mooney. Semi-supervised clustering by seeding. In *International Conference on Machine Learning (ICML)*, pages 19–26, 2002.

[4] S. Basu, M. Bilenko, and R. Mooney. A probabilistic framework for semi-supervised clustering. In *10th ACM SIGKDD*, pages 59–68, 2004.

[5] F. Beil, M. Ester, and X. Xu. Frequent term-based text clustering. In *8th ACM SIGKDD*, pages 436–442, 2002.

[6] H. Cheng, K. Hua, and K. Vu. Constrained locally weighted clustering. *Proceedings of the VLDB Endowment*, 1(1):90–101, 2008.

[7] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[8] B. Fung. Hierarchical document clustering using frequent itemsets. In *3rd SIAM International Conference on Data Mining (SDM)*, pages 59–70, 2003.

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

[10] A. Huang, D. Milne, E. Frank, and I. Witten. Clustering documents with active learning using wikipedia. In *Int. Conf. on Data Mining (ICDM)*, pages 839–844, 2008.

[11] R. Huang and W. Lam. An active learning framework for semi-supervised document clustering with language modeling. *Data & Knowl. Eng.*, 68(1):49–67, 2009.

[12] X. Ji and W. Xu. Document clustering with prior knowledge. In *29th ACM SIGIR Conf. on Research and Development in Inf. Retrieval*, pages 405–412, 2006.

[13] B. Liu, X. Li, W. S. Lee, and P. S. Yu. Text classification by labeling words. In *19th national conference on Artifical intelligence*, AAAI'04, pages 425–430, 2004.

[14] H. Malik and J. Kender. High quality, efficient hierarchical document clustering using closed interesting itemsets. In *6th International Conference on Data Mining (ICDM)*, pages 991–996, 2006.

[15] B. Nogueira, M. Moura, M. Conrado, R. Rossi, R. Marcacini, and S. Rezende. Winning some of the document preprocessing challenges in a text mining process. In *IV Workshop on Algorithms and Data Mining Applications*, pages 10–18, 2008.

[16] L. Rigutini and M. Maggini. A semi-supervised document clustering algorithm based on EM. In *International Conference on Web Intelligence (IEEE/WIC/ACM)*, pages 200–206, 2005.

[17] P. Tan, M. Steinbach, V. Kumar, et al. *Introduction to data mining*. Pearson Addison Wesley Boston, 2006.

[18] W. Zhang, T. Yoshida, X. Tang, and Q. Wang. Text clustering using frequent itemsets. *Knowledge-Based Systems*, 23(5):379–388, 2010.