Filtering Methods for Feature Selection in Web-Document Clustering

Heum Park and Hyuk-Chul Kwon

AI Lab. Dept. of Computer Science, Pusan National University, Busan, Korea
parkheum2@empal.com, hckwon@pusan.ac.kr

Abstract. This paper presents the results of a comparative study of filtering methods for feature selection in web document clustering. First, we focused on feature selection methods based on Mutual Information (MI) and Information Gain (IG). With those features and feature values, and using MI and IG, we extracted from documents representative max-value features as well as a representative cluster for a feature and a representative cluster for a document. Second, we tested the Max Feature Selection Method (MFSM) with those representative features and clusters, and evaluated the web-document clustering performance. However, when document sets yield poor clustering results by term frequency, we cannot obtain good features using the MFSM with the MI and IG values. Therefore, we propose new filtering methods, Min Count of Representative Cluster for a Feature (MCRCF) and Min Count of Representative Cluster for a Document (MCRCD). In the experimental results, the MFSM showed better performance than was achieved using only term frequency, MI and IG. And when we applied the new filtering methods for feature selection (MCRCF, MCRCD), the clustering performance improved notably. Thus we can assert that those filtering methods are effective means of feature selection and offer good performance in web document clustering.

Keywords: feature selection, feature filtering, document clustering.

1 Introduction

Among feature selection methods in web-document clustering, X^2 , Mutual Information(MI), and Information Gain(IG) are widely used.[1][2] However, we cannot rely on such methods to provide consistently good performance, because web documents have either a small number of terms or many kinds of terms and high term frequencies. A small number of terms can be maintained in a document by increasing the weights of feature values, but for documents containing many kinds of terms and high term frequencies, the document vector space has to be reduced using filtering methods. Filtering methods are general preprocessing algorithms that do not rely on any knowledge of the algorithm.[3][4][5] Therefore, we propose new feature filtering methods that can be applied independently to clustering application programs.

In the present study, first we extracted features and feature values using MI and IG. And with those features and values, we extracted the documents' representative

max-value features as well as a representative cluster for a feature and a representative cluster for a document. Second, we used the Max Feature Selection Method (MFSM), which selects features using the max-feature values and a representative cluster number for a feature. And we proposed Min Count of Representative Cluster for a Feature (MCRCF) method and Min Count of Representative Cluster for a Document (MCRCD) methods for filtering features.

2 Preliminaries

We used three corpora, 'the Natural Science' directory services at www.empas.com, www.yahoo.co.kr, and www.naver.com, three famous Korean portal sites, as experimental datasets for feature selection and clustering. Those documents are well classified manually by an indexer, and so we could easily evaluate the clustering performance by comparing the pre-allocated directory with the clustering results. Accordingly, we selected 9 subdirectories within the directory services. The total number of document sets was 1,036, 964 and 1,093 respectively, and the term vector space was a matrix consisting of tf or $tf*idf$. After bisecting each document set, we used one half for feature selection and clustered the other half by applying the selected features. For clustering and analysis of clustering performance, we used a clustering toolkit, Cluto2.1, developed by the University of Minnesota.[3] We compared the results of clustering using entropy- and purity- averaged measures.

For feature selection, first we calculated the feature values using the MI and IG, and with the features and feature values, allocated each term to only the one category c that had the maximum score for each expression $MI_{max} = \max\{MI(t, c_i)\}$, and $(IG_h)_{max} = \max\{IG(t, c_i)\}$, { if $avg(IG(t, c_i)) > \Gamma$, (Γ : threshold)}, and applied those values in a document matrix. Second, with the feature values $FV_c(t)$, we obtained representative clusters for a feature and representative clusters for a document. The representative cluster number $C_f(t) = \max_{i=1}^k \{FV_i(t)\}$ for a term t is the cluster number

with the maximum value among all of the clusters(i is the cluster number). And third, we obtained the representative cluster number $C_f(d) = \max_{i=1}^k \{Count_i(C_i(t_1 : t_n))\}$

for each document using $C_f(t)$, where $Count_i(C_i(t_1 : t_n))$ is the number of terms $t_1 \dots t_n$ that have $C_f(t)$ in the i -th cluster for a document.

3 Filtering Methods for Feature Selection

First, we used the MFSM to filter features. We calculated the max-feature values $FV_i(t)_{max}$ in the i -th cluster using MI and IG, and with $FV_i(t)_{max}$, we determined the representative cluster number $C_f(t)$ for a term t . We then selected features in a representative cluster. In expression 1, $count(C_i(t))$ is the count $C_i(t)$ for a term t in all of the clusters. Second, with MCRCF, we selected features using $FV_i(t)_{max}$ and $C_f(t)$ and thresholds. If the feature value $FV_i(t)$ in the i -th cluster for a term t is greater than the threshold λ , and the count of $C_i(t)$ of term t in all of the clusters is greater than the threshold T (the threshold is 50% or 40% or 30% of the total number of clusters), those features are removed from a document matrix. For example, when term t has a

feature value $FV_i(t)_{max}$ greater than the threshold λ ($\lambda=0$) in cluster 1 and cluster 3, $count(C_i(t))$ is 2, and if the threshold T is 3, term t is selected. Third, with MCRCD, we calculated the $FV_i(t)_{max}$ and $C_i(t)$ for a term t in the i -th cluster, as with MCRCF. We could obtain a representative cluster number for a document $C_f(d)$ as in section 2. To apply the MCRCD method, we require the sum of feature values for term t in all of the clusters, the weight W (2 or 3 or 4) and thresholds T (50% or 40% or 30%) constants. We include the following 5 steps in the filtering process.

Step 1: Calculate the feature value $FV(t)$ of term t and $FV_i(t)_{max}$ in the i -th cluster.

Step 2: Obtain $C_i(t)$ and $C_i(d)$ using $FV_i(t)_{max}$ in the i -th cluster.

Step 3: Find features F_i that have a feature value $FV(C_f(d))$ greater than $(1/W) * \sum FV_i(C_i(d))$ and λ , as in expression 2, where W is the weight constant. Repeat Step3 for all of the terms from a document.

Step 4: Using those features F_i obtained from Step3, as in expression 3, if the count of occurrences $C_i(F_i)$ of feature F_i in all of the clusters is greater than the threshold T , those features are removed. Repeat Step2 to Step3 for all of the documents.

Step 5: Apply those features and feature values to a document matrix, and cluster.

$$count_{i=1}^k (C_i(t)) < Threshold(T) \quad (1)$$

$$\left(\frac{1}{W}\right) * \sum_{i=1}^k FV_i(C_i(d)) < FV(C_f(d)), \text{ and } FV() > \lambda \quad (2)$$

$$count_{i=1}^k (C_i(F_i)) < Threshold(T) \quad (3)$$

Fig. 1. Entropies and Purities averaged using TF, MI, IG, and Filtering Methods MFSM with MI and IG, MCRCF, and MCRCD. The clustering number is 30.

4 Experimental Evaluations

In these experiments, we clustered the document matrix by applying the MFSM with the MI and IG values. As shown in Fig 1, the entropies and purities of *Empas(1)* were the results of clustering the document set *Empas(1)* by applying features selected by

Empas(2). In the case of the *Empas* and *Yahoo* document sets, the clustering results for the MFSM with the MI and IG values (MFSM-MI, MFSM-IG) were better than for TF, MI and IG. But in the case of the *Naver* document sets, the clustering results for MI and IG were very poor. Because the results by TF had poor entropies and purities (0.327, 0.731), we could not obtain good features using the MFSM with MI and IG values. And when we used the MCRCF and MCRCF methods, for not only the *Empas* and *Yahoo* document sets but also for the *Naver* document sets, the clustering results were better than when using the MFSM with MI and IG values, in all of the entropies and purities. The entropies and purities of the clustering results using MCRCF and MCRCF were improved notably.

5 Conclusion

This paper presents the results of a comparative study on features filtering for feature selection web-document clustering. We applied, as feature selection methods, MI and IG, and then the MFSM. There were better results for some document sets than those obtained using term frequency (TF). But because the results by TF yielded poor entropies and purities, we could not obtain good features using MI and IG, or the MFSM either. Therefore, we were obliged to find new methods for selecting good features and achieving good performance. When we applied MCRCF and MCRCF, we were able to obtain a much better performance than that achieved using the MFSM with MI and IG. Most notably in the case of the *Yahoo* document sets, there was an extraordinarily good performance for all data sets using MCRCF and MCRCF. Therefore, we can confirm that these feature-filtering methods offer enhanced clustering performance as well as an effective means of selective filtering.

Acknowledgments. “This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD)” (The Regional Research Universities Program/Research Center for Logistics Information Technology).

References

1. Yiming Yang and Jan O. Pederson : A comparative study on feature selection in text categorization, Proceedings of ICML-97, 14th International Conference on Machine Learning, (1997).
2. Gang Wang, Frederick H. Lochovsky : Feature selection with conditional mutual information maximin in text categorization. Proceedings of the thirteenth ACM international conference, (2004), 342-349.
3. Ying Zhao and George Karypis : Criterion Functions for Document Clustering: Experiments and Analysis. Technical Report TR #01--40, Department of Computer Science (2002)
4. Sanmay Das : Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection, The Proceedings of the Eighteenth International Conference on Machine, (2001) 74 – 81
5. Huang Yuan, Shian-Shyong Tseng, Wu Gangshan, Zhang Fuyan : A two-phase feature selection method using both filter and wrapper, IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference (1999) vol.2 132-136,
6. Baranidharan Raman, Thomas R. Iroger : Instance Based Filter for Feature Selection, journal of Machine Learning Research 1, (2002) 1-23,.