Abstract

A powerful technique that has been widely used to organizing a large number of web documents into a small number of general and meaningful clusters is Document Clustering. High dimensionality, scalability, accuracy, extracting semantics relations from texts and meaningful cluster labels are the major challenges for document clustering. To improve the document clustering quality, we intend to introduce an effective methodological system using association rules instead of frequent term sets for clustering web documents into different topical groups called Hard Document Clustering using Association Rules (HDCAR). HDCAR characterized by high performance in the organization of web documents and navigates them effectively in order to keep up with the explosive growth of the number and size of web documents. Association Rule has the equally important advantage of having a higher descriptive power compared to single words (frequent term sets). Moreover, the external knowledge from both WordNet synonym and hypernyms will be used to enhance the “bag of words” used before the clustering process and to assist the label generation procedure following the clustering process. Then, Multi-Hash Tire Association Rule (MHTAR) algorithm is
used to discover a set of highly-related association rules to overcome the drawbacks of the Apriori algorithm. Through the resulted association rules, the hidden topics are discovered as the first step and then the documents will be cluster based on them. Finally, each document is assigned to only one cluster (hard clustering) with the highest Document Weighted-measure, and then the highly similar clusters are merged. To evaluate the performance of HDCAR, we conducted experiments based on four different kinds of datasets Classic, Re0, WebKB and REUTER datasets. The experimental results show that HDCAR outperforms the major document clustering methods like k-means, Bisecting k-means, FIHC, and UPGMA with higher accuracy quality, efficiency and lower execution time. Furthermore, HDCAR provides more general and meaningful labels for documents and increases the documents clusterization speed, as a result of the reduction of their dimensionality.

References

15. L. Parsons, E. Haque, H..Liu, “Subspace clustering for high dimensional data: A review”,
Investigate the Performance of WordNet and Association Rules for Hard Clustering Web Document


46.

**Index Terms**

Computer Science  
Information Sciences

**Keywords**

Web Mining, Document Clustering, Association rule mining, WordNet, Fuzzy weighting score.