ESPClust: An Effective Skew Prevention Method for Model-based Document Clustering[*]

Xiaoguang Li, Ge Yu, Daling Wang, Yubin Bao

School of Information Science and Engineering, Northeastern University
Shenyang 110004, P.R.China
{xgli7312@mail.163.com yuge@mail.neu.edu.cn}

Abstract. Document clustering is necessary for information retrieval, Web data mining, and Web data management. To support very high dimensionality and the sparsity of document feature, the model-based clustering has been proved to be an intuitive choice for document clustering. However, the current model-based algorithms are prone to generating the skewed clusters, which influence the quality of clustering seriously. In this paper, the reasons of skew generating are examined and determined as the inappropriate initial model, and the interaction between the decentralization of estimation samples and the over-generalized cluster model. An effective clustering skew prevention method (ESPClust) is proposed to focus on the last reason. To break this interaction, for each cluster, ESPClust automatically selects a part of documents that most relevant to its corresponding class as the estimation samples to re-estimate the cluster model. Based on the ESPClust, two algorithms with respect to the quality and efficiency are provided for different kinds of applications. Compared with balanced model-based algorithms, the ESPClust method has less restrictions and more applicability. The experiments show that the ESPClust can avoid the clustering skew in a great degree and its Macro-F1 measure outperforms the previous methods' measure.

1 Introduction

In recent years, there is a tremendous growth in the volume of documents available on Web, digital libraries, and news media. This has led to an increased interest in developing methods that can help users to effectively navigate, summarize, and organize this information or to discovery the inherent knowledge underlying document collection. As an important technique towards these goals, a high-quality document clustering algorithm plays an important role for information retrieval, Web data mining, and Web data management.

[*] Supported by the National Natural Science Foundation of China under Grant No.60173051 and the Teaching and Research Award Program for Outstanding Young Teachers in Higher Education Institution of the Ministry of Education, China