Chapter XI
Some Efficient and Fast Approaches to Document Clustering

P. Viswanth
Indian Institute of Technology Guwahati, India

Bidyut Kr. Patra
Indian Institute of Technology Guwahati, India

V. Suresh Babu
Indian Institute of Technology Guwahati, India

ABSTRACT

Clustering is a process of finding natural grouping present in a dataset. Various clustering methods are proposed to work with various types of data. The quality of the solution as well as the time taken to derive the solution is important when dealing with large datasets like that in a typical documents database. Recently hybrid and ensemble based clustering methods are shown to yield better results than conventional methods. The chapter proposes two clustering methods; one is based on a hybrid scheme and the other based on an ensemble scheme. Both of these are experimentally verified and are shown to yield better and faster results.

INTRODUCTION

Document clustering is to cluster a set of text documents. The clustering result which reveals the structure present in the data can be used for improving an information retrieval system (Gerald Kowalski, 1997 and Tombros, Villa & Van Rijsbergen, 2002), to present the results of a search engine in a more

structured way (Zamir, Etzioni, & Madani, 1997), *etc.* Document clustering can be seen as a special case of clustering and has several unique problems related to representation schemes for documents, similarity measures to compare documents, clustering methods, evaluation of the results, *etc.*

Various clustering methods are proposed which can be categorized as hierarchical methods, partition based methods, density based methods, *etc*., based on the way the method works. Each method in each of these categories has its own advantages and disadvantages. For example, hierarchical methods like *single-link* method (Jain, Murty & Flynn, 1999) and density based methods like *DBSCAN* (Ester, Kriegel & Xu, 1996) are more time consuming than partition based methods like *k-means*. But the results of hierarchical or density based methods are in general better than that of partition based methods. *Single-link* method is sensitive to noise whereas *DBSCAN* can find noisy outliers. The time complexity of the clustering method is very important where the dataset sizes are very large, and typically this is the case with document databases. Also the quality of a result is important. To meet these conflicting ends, recently *hybrid* and *ensemble* based methods are proposed.

Hybrid clustering methods combine two or more clustering methods to achieve a better result (Surdeanu, Turmo, & Ageno, 2005 and Peng, Kou, Shi, & Chen, 2006) and or to derive a quicker result (Lin, & Chen, 2005 and Viswanath, & Rajwala, 2006). Typically, applying initially a less time consuming method like a partition based method which yields a smaller set of prototypes which in turn are used with a critical hierarchical or density based methods are promising. It can yield results of high quality which consumes less time.

Ensemble based clustering methods (Viswanath, & Jayasurya, 2006) are a recent research advancement motivated from ensemble based classifiers. Ensemble of classifiers is a relatively well studied area than ensemble of clustering methods. Ensemble of classifiers where the classification decisions from multiple classifiers are combined is shown to improve the performance (Fumera & Roli, 2005). The interesting observation is that even though the individual component classifiers are weak methods (they need to be only slightly superior than random guessing) their ensemble can be a quite stronger one (Freund & Schapire, 1999). Recently there are some studies on ensemble of clustering methods applied with document data which basically combines various partitions of the data set obtained from component clustering methods into a single partition (Greene & Cunningham, 2006).

The chapter, after discussing some relevant issues with hybrid and ensemble based clustering methods, presents two fast and efficient clustering methods where one is a hybrid method and the other is an ensemble method. Both of these methods run in a linear time of the input data size, but it can find an on-par clustering results as that of density based and hierarchical methods.

BACKGROUND

Clustering is a process of finding *groups* called *clusters* present in a given data set such that the data items present in a cluster are similar to each other, whereas those present in different clusters are dissimilar. There are various clustering methods applied in various fields which use various similarity measures (Jain, Murty & Flynn, 1999). Even though the problem seems simple and a relatively older one, it is still an active research area, and recently it is shown that there is no clustering method which satisfies certain simple properties (Kleinberg, 2002). A good clustering method in one field need not be a good one in some other field.