# Hierarchical Document Clustering

**Benjamin C. M. Fung**
*Simon Fraser University, Canada*

**Ke Wang**
*Simon Fraser University, Canada*

**Martin Ester**
*Simon Fraser University, Canada*

## INTRODUCTION

Document clustering is an automatic grouping of text documents into clusters so that documents within a cluster have high similarity in comparison to one another, but are dissimilar to documents in other clusters. Unlike document classification (Wang, Zhou, & He, 2001), no labeled documents are provided in clustering; hence, clustering is also known as unsupervised learning. Hierarchical document clustering organizes clusters into a tree or a hierarchy that facilitates browsing. The parent-child relationship among the nodes in the tree can be viewed as a topic-subtopic relationship in a subject hierarchy such as the Yahoo! directory.

This chapter discusses several special challenges in hierarchical document clustering: high dimensionality, high volume of data, ease of browsing, and meaningful cluster labels. State-of-the-art document clustering algorithms are reviewed: the partitioning method (Steinbach, Karypis, & Kumar, 2000), agglomerative and divisive hierarchical clustering (Kaufman & Rousseeuw, 1990), and frequent itemset-based hierarchical clustering (Fung, Wang, & Ester, 2003). The last one, which was recently developed by the authors, is further elaborated since it has been specially designed to address the hierarchical document clustering problem.

## BACKGROUND

Document clustering is widely applicable in areas such as search engines, web mining, information retrieval, and topological analysis. Most document clustering methods perform several preprocessing steps including stop words removal and stemming on the document set. Each document is represented by a vector of frequencies of remaining terms within the document. Some document clustering algorithms employ an extra preprocessing step that divides the actual term frequency by the overall frequency of the term in the entire document set. The idea is that if a term is too common across different documents, it has little discriminating power (Rijsbergen, 1979). Although many clustering algorithms have been proposed in the literature, most of them do not satisfy the special requirements for clustering documents:

- **High Dimensionality:** The number of relevant terms in a document set is typically in the order of thousands, if not tens of thousands. Each of these terms constitutes a dimension in a document vector. Natural clusters usually do not exist in the full dimensional space, but in the subspace formed by a set of correlated dimensions. Locating clusters in subspaces can be challenging.
- **Scalability:** Real world data sets may contain hundreds of thousands of documents. Many clustering algorithms work fine on small data sets, but fail to handle large data sets efficiently.
- **Accuracy:** A good clustering solution should have high intra-cluster similarity and low inter-cluster similarity, i.e., documents within the same cluster should be similar but are dissimilar to documents in other clusters. An external evaluation method, the F-measure (Rijsbergen, 1979), is commonly used for examining the accuracy of a clustering algorithm.
- **Easy to Browse with Meaningful Cluster Description:** The resulting topic hierarchy should provide a sensible structure, together with meaningful cluster descriptions, to support interactive browsing.
- **Prior Domain Knowledge:** Many clustering algorithms require the user to specify some input parameters, e.g., the number of clusters. However, the user often does not have such prior domain knowledge. Clustering accuracy may degrade drastically if an algorithm is too sensitive to these input parameters.

## DOCUMENT CLUSTERING METHODS

### Hierarchical Clustering Methods

One popular approach in document clustering is agglomerative hierarchical clustering (Kaufman & Rousseeuw, 1990). Algorithms in this family build the hierarchy bottom-up by iteratively computing the similarity between all pairs of clusters and then merging the most similar pair. Different variations may employ different similarity measuring schemes (Karypis, 2003; Zhao & Karypis, 2001). Steinbach (2000) shows that Unweighted Pair Group Method with Arithmatic Mean (UPGMA) (Kaufman & Rousseeuw, 1990) is the most accurate one in its category. The hierarchy can also be built top-down which is known as the divisive approach. It starts with all the data objects in the same cluster and iteratively splits a cluster into smaller clusters until a certain termination condition is fulfilled.

Methods in this category usually suffer from their inability to perform adjustment once a merge or split has been performed. This inflexibility often lowers the clustering accuracy. Furthermore, due to the complexity of computing the similarity between every pair of clusters, UPGMA is not scalable for handling large data sets in document clustering as experimentally demonstrated in (Fung, Wang, & Ester, 2003).

### Partitioning Clustering Methods

K-means and its variants (Cutting, Karger, Pedersen, & Tukey, 1992; Kaufman & Rousseeuw, 1990; Larsen & Aone, 1999) represent the category of partitioning clustering algorithms that create a flat, non-hierarchical clustering consisting of k clusters. The k-means algorithm iteratively refines a randomly chosen set of k initial centroids, minimizing the average distance (i.e., maximizing the similarity) of documents to their closest (most similar) centroid. The bisecting k-means algorithm first selects a cluster to split, and then employs basic k-means to create two sub-clusters, repeating these two steps until the desired number k of clusters is reached. Steinbach (2000) shows that the bisecting k-means algorithm outperforms basic k-means as well as agglomerative hierarchical clustering in terms of accuracy and efficiency (Zhao & Karypis, 2002).

Both the basic and the bisecting k-means algorithms are relatively efficient and scalable, and their complexity is linear to the number of documents. As they are easy to implement, they are widely used in different clustering applications. A major disadvantage of k-means, however, is that an incorrect estimation of the input parameter, the number of clusters, may lead to poor clustering accuracy. Also, the k-means algorithm is not suitable for discovering clusters of largely varying sizes, a common scenario in document clustering. Furthermore, it is sensitive to noise that may have a significant influence on the cluster centroid, which in turn lowers the clustering accuracy. The k-medoids algorithm (Kaufman & Rousseeuw, 1990; Krishnapuram, Joshi, & Yi, 1999) was proposed to address the noise problem, but this algorithm is computationally much more expensive and does not scale well to large document sets.

### Frequent Itemset-Based Methods

Wang et al. (1999) introduced a new criterion for clustering transactions using frequent itemsets. The intuition of this criterion is that many frequent items should be shared within a cluster while different clusters should have more or less different frequent items. By treating a document as a transaction and a term as an item, this method can be applied to document clustering; however, the method does not create a hierarchy of clusters.

The Hierarchical Frequent Term-based Clustering (HFTC) method proposed by (Beil, Ester, & Xu, 2002) attempts to address the special requirements in document clustering using the notion of frequent itemsets. HFTC greedily selects the next frequent itemset, which represents the next cluster, minimizing the overlap of clusters in terms of shared documents. The clustering result depends on the order of selected itemsets, which in turn depends on the greedy heuristic used. Although HFTC is comparable to bisecting k-means in terms of clustering accuracy, experiments show that HFTC is not scalable (Fung, Wang, & Ester, 2003).

### A Scalable Algorithm for Hierarchical Document Clustering: FIHC

A scalable document clustering algorithm, Frequent Itemset-based Hierarchical Clustering (FIHC) (Fung, Wang, & Ester, 2003), is discussed in greater detail because this method satisfies all of the requirements of document clustering mentioned above. We use “item” and “term” as synonyms below. In classical hierarchical and partitioning methods, the pairwise similarity between documents plays a central role in constructing a cluster; hence, those methods are “document-centered”. FIHC is “cluster-centered” in that it measures the cohesiveness of a cluster directly using frequent itemsets: documents in the same cluster are expected to share more common itemsets than those in different clusters.

A frequent itemset is a set of terms that occur together in some minimum fraction of documents. To illustrate the usefulness of this notion for the task of clustering, let us consider two frequent items, “win-