# Preceding Document Clustering by Graph Mining Based Maximal Frequent Termsets Preservation

Syed Shah and Mohammad Amjad
Department of Computer Engineering, Jamia Millia Islamia, India

**Abstract:** *This paper presents an approach to cluster documents. It introduces a novel graph mining based algorithm to find frequent termsets present in a document set. The document set is initially mapped onto a bipartite graph. Based on the results of our algorithm, the document set is modified to reduce its dimensionality. Then, Bisecting K-means algorithm is executed over the modified document set to obtain a set of very meaningful clusters. It has been shown that the proposed approach, Clustering preceded by Graph Mining based Maximal Frequent Termsets Preservation (CGFTP), produces better quality clusters than produced by some classical document clustering algorithm(s). It has also been shown that the produced clusters are easily interpretable. The quality of clusters has been measured in terms of their F-measure.*

**Keywords:** *Bipartite graph, graph mining, frequent termsets mining, bisecting K-means.*

*Received June 18, 2016; accepted June 29, 2017*

## 1. Introduction

Document clustering (or text clustering) is the application of cluster analysis to textual documents. It has applications in text mining, automatic document organization, topic extraction and fast information retrieval or filtering. The applications may be online or offline. Originally document clustering was used to improve the precision in information retrieval systems [10, 16] and for finding the nearest neighbors of a document [2]. Later it was found to be useful in browsing text documents (e.g., news articles) [19] and also for organizing the results of a web user's query onto a search engine [22]. It has also been used to generate hierarchical clusters of documents [9].

Two well known document clustering techniques are K-means and agglomerative hierarchical clustering. K-means is faster than agglomerative hierarchical clustering (hierarchical clustering has a quadratic time complexity in contrast to a variant of K-means that has a linear time complexity) [18]. But both these algorithms do not address the fundamental problem of document clustering that of very high dimensionality.

Later a new approach was introduced where a clustering algorithm was not applied to the input document corpus but instead to its refined form that constitutes not all terms but only those that are frequent. Thus two fields of data mining-frequent termsets mining and clustering - were treated as two phases of document clustering, one after another. Based on this new approach many algorithms were proposed (see section 1.1.).

### 1.1. Related Work

Morzy *et al*. [15] introduces a hierarchical clustering algorithm that uses sequential patterns found in the database to generate both the clustering model and data clusters, [13] in one approach Clustering based on Frequent Word Sequences (CFWS) takes into consideration the sequence of frequent words where {cricket, bat} and {bat, cricket} are treated as two different patterns and in another approach Clustering based on Frequent Word Meaning Sequences (CFWMS) takes into consideration frequent word meaning sequences where not only sequence but also the contextual meaning of each word has been considered, [11] applies frequent-itemset based clustering to web search results, [6] proposes document clustering based on maximal frequent sequences where only maximal word sequences have been considered, [1] starts with an empty set, it continues selecting one more element (one cluster description) from the set of remaining frequent itemsets until the entire document collection is contained in the cover of the set of all chosen frequent itemsets, [4] takes into consideration both global frequent items and cluster frequent items, [23] finds frequent itemsets and then uses minimum spanning tree algorithm to construct clusters, [21] groups web transactions using a hierarchical pattern-based clustering approach, [12] uses Apriori for finding frequent itemsets and then uses the mined frequent itemsets to obtain partitions (with no overlapping) and after that groups documents within a partition using derived keywords, some other researches propose use of Wikipedia as external knowledge source thus taking into consideration the semantic relationships between words [8], for dynamic data [17] proposed evolutionary clustering that was again based on frequent itemsets, to decrease the number of patterns and time complexity [14] proposed a pattern-based hierarchical document clustering that mines for local patterns and builds a cluster hierarchy