N-Clustering of Text Documents Using Graph Mining Techniques

Bapuji Rao
https://orcid.org/0000-0002-2781-9708
Indira Gandhi Institute of Technology, Sarang, India

INTRODUCTION

Text mining refers to the detection of patterns or similarities in natural language text. For given group of text documents, often the need arises to classify the documents into groups or clusters based on similarity of content. For a small collection, it is possible to manually perform the grouping of documents into specific groups. But to group large volumes of text, the process would be extremely more consumption of time. Therefore, developing a fast and accurate document clustering algorithm is very much helpful for processing of document data sets for clustering.

Clustering is generally studied in data mining problem of text domains. Clustering is an automatic learning technique which aims at grouping a set of objects into Sub-sets or clusters. A clustering can be defined as a grouping of similar objects in the data of similar characteristics. The similarity of characteristics between the objects is measured with the help of a similarity function. The problem of clustering can be very useful in the text domain. Document clustering is extensively used in the areas of text mining and information retrieval. Clustering especially helps of organizing documents in a structural way to improve retrieval and browsing those documents. The study of the clustering problem is related to the applicability to the text domain. Text document clustering is a selection of text documents with the particular word(s)/text(s) present. So each group of text documents called a cluster of text documents of a particular word's presence. Clustering is unsupervised learning it means there is no need of human interference for clustering of documents. In text document clustering, a group of words (texts) is used on a set of text documents for discovering such text documents having with the given set of words (texts). Further such discovered text documents for the given set of words (texts) are grouped into that many clusters of text documents.

Graph mining, is also known as graph-based data mining, is the extraction of meaningful and useful knowledge from a graph representation of data. The most natural form of knowledge that can be extracted from graphs is also a graph. Therefore, the knowledge, sometimes referred to as patterns (or sequence), mined from the data are typically expressed as graphs, which may be sub-graphs of the graphical data.

A Graph G = (V, E) with nodes or vertices in a set of vertices V and a set of edges E. Two nodes u, v Î V are adjacent, if an edge (u, v) Î E exists. The graph representation, that is, a collection of nodes and links between nodes, does support all aspects of the relational data mining process. As one of the most general forms of data representation, the graph easily represents entities, their attributes, and their relationships to other entities. In a bi-partite graph, there are two types of nodes are used. In document clustering, the two types of nodes are document and word. Between the documents and words, the formation of graph is known as document-word bi-partite graph. The relationship between the document-word is the frequency of words present in the related document.

DOI: 10.4018/978-1-7998-3479-3.ch057

BACKGROUND

Document Clustering is also known as Text Clustering is a specific application of text mining and a sub-problem of cluster analyses. If the documents of certain categories are identified and grouped as one cluster, such is called Text Classification Problem. The approach discussed in this chapter is applicable to text documents having characteristics of common word(s) appearances in those text documents for clustering purpose. Finally, each cluster of text documents is represented as a bi-partite graph. In a bi-partite graph, the two types of nodes are document node and word node.

LITERATURE SURVEY

The Scatter-Gather method proposed by authors (Cutting, Karger, Pedersen, & Tukey, 1992) defines the hierarchical organization of documents into coherent categories for systematic browsing of the document collection. It provides a systematic browsing technique with the use of clustered organization of the document collection.

The authors (Aggarwal & Zhai, 2012) define both feature selection and feature transformation methods such as Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Analysis (PLSA), and Non-negative Matrix Factorization (NMF) are used to improve the quality of the document representation and make it more efficient to text clustering. Feature selection is more common and easy to apply in text clustering in which supervision is available for the feature selection process proposed by authors (Yang & Pedersen, 1997). Since the results of text clustering are highly dependent on document similarity. Such cases the concept of term contributed by authors (Liu, Liu, Chen, & Ma, 2003) is applied. So the contribution of a term can be viewed as its contribution to document similarity.

The technique of concept decomposition uses any standard clustering technique has been studied by authors (Aggarwal & Yu, 2001); (Dhillon, & Modha, 2001) on the original representation of the documents. The frequent terms in the centroids of these clusters are used as basis vectors which are almost orthogonal to one another. The documents can then be represented in a much more concise way in terms of these basis vectors. So the condensed conceptual representation allows for enhanced clustering as well as classification of text documents. Therefore, a second phase of clustering can be applied on this condensed representation in order to cluster the documents much more effectively proposed by author (Salton, 1983). Such a method is tested by authors (Slonim & Tishby, 2000) by using word-clusters in order to represent documents.

The non-negative matrix factorization (NMF) technique is a latent space method, and particularly suitable for clustering of text documents proposed by (Xu, Liu, & Gong, 2003). The NMF scheme is a feature transformation method which is particularly used for clustering of documents. Let A be the $n \times d$ term document matrix. To create k clusters from the underlying documents. Then, the non-negative matrix factorization method attempts to determine the matrices U and V by minimizing the following objective function: $J = (1/2) * \| A\text{-}U*V^T\|$. The matrix factorization technique is used to determine word clusters instead of document clusters. Just as the columns of V provide a basis which can be used to discover document clusters, just like the use of columns of U to discover a basis which correspond to word clusters. Since document clusters and word clusters are closely related, and it is often useful to discover both simultaneously, and considered as co-clustering. The co-clustering is proposed by (Dhillon, 2001); (Dhillon & Modha, 2001); (Dhillon, Mallela, & Modha, 2003).