An Approach to Clustering of Text Documents Using Graph Mining Techniques

Bapuji Rao, BPUT, Rourkela, India
Brojo Kishore Mishra, Department of Information Technology, C. V. Raman College of Engineering, Bhubaneswar, India

ABSTRACT

This paper introduces a new approach of clustering of text documents based on a set of words using graph mining techniques. The proposed approach clusters (groups) those text documents having searched successfully for the given set of words from a set of given text documents. The document-word relation can be represented as a bi-partite graph. All the clustering of text documents is represented as sub-graphs. Further, the paper proposes an algorithm for clustering of text documents for a given set of words. It is an automated system and requires minimal human interaction for the clustering of text documents. The algorithm has been implemented using C++ programming language and observed satisfactory results.

KEYWORDS

Bi-partite Graph Clustering, Self-loop, Sub-graph, Weighted Un-Oriented Incidence Matrix

INTRODUCTION

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 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 has been extensively used in a number of different 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 precedes its applicability to the text domain. Text document clustering is a selection of text documents with the particular word(s) present. So each group of text documents called cluster of text documents of a particular word’s presence. Clustering is the most common form of unsupervised learning and no supervision means that there is no human expert who has assigned documents to cluster. In text document clustering, it is the distribution and makeup of the text documents as a group based on a particular word present in all the grouped text documents. Clustering is sometimes referred to as automatic classification. In text document clustering, a group of words are used on a collection of text documents for discovering such text documents having with the given set of words. Further such discovered text documents for the given set of words are grouped into that many cluster of text documents.
LITERATURE REVIEW

The Scatter-Gather method in (Cutting, Karger, Pedersen, & Tukey, 1992) says 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.

In the article by (Aggarwal & Zhai, 2012), the author says 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 (Yang & Pedersen, 1997). Since the results of text clustering are highly dependent on document similarity. Such cases the concept of term contributed by (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 in past studies (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 by (Salton, 1983). Such a method is tested in (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×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-U*V||. 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 found in (Dhillon, 2001); (Dhillon & Modha, 2003); (Dhillon, Mallela, & Modha, 2003).

Distance-based clustering algorithms are designed by using a similarity function to measure the closeness between the text objects. The most well-known similarity function which is used commonly in the text domain is the cosine similarity function proposed by (Aggarwal & Zhai, 2012). These similarity functions can be used in conjunction with a wide variety of traditional clustering algorithms has been studied in past studies (Jain & Dubes, 1998); (Kaufman, & Rousseeuw, 2009).

The k-means clustering algorithm also uses a set of k representatives around which the clusters are built. The simplest form of the k-means approach is to start with a set of k seeds from the original set of documents, and assign documents to these seeds on the basis of closest similarity. In the next iteration, the centroid of the assigned points to each seed is used to replace the seed in the last iteration was proposed by (Aggarwal & Zhai, 2012).

A graph-based approach to document classification is described. The document sets are represented as graph sets where weighted graph mining algorithm is applied to extract frequent sub-graphs. Further the sub-graphs are processed to produce feature arrays (one per document) for classification. Finally, the
Crisis Compliance: Using Information Technology to Predict, Prevent and Prevail over Disasters
Laura Lally (2010). *Breakthrough Discoveries in Information Technology Research: Advancing Trends* (pp. 137-150).
[www.igi-global.com/chapter/crisis-compliance-using-information-technology/39576?camid=4v1a](www.igi-global.com/chapter/crisis-compliance-using-information-technology/39576?camid=4v1a)

The Serious Games Applied for Health
[www.igi-global.com/chapter/the-serious-games-applied-for-health/113018?camid=4v1a](www.igi-global.com/chapter/the-serious-games-applied-for-health/113018?camid=4v1a)