Chapter 12
A Brief Review of Metaheuristics for Document or Text Clustering

Sinem Büyüksaatçı
Istanbul University, Turkey

Alp Baray
Istanbul University, Turkey

ABSTRACT

Document clustering, which involves concepts from the fields of information retrieval, automatic topic extraction, natural language processing, and machine learning, is one of the most popular research areas in data mining. Due to the large amount of information in electronic form, fast and high-quality cluster analysis plays an important role in helping users to effectively navigate, summarize and organise this information for useful data. There are a number of techniques in the literature, which efficiently provide solutions for document clustering. However, during the last decade, researchers started to use metaheuristic algorithms for the document clustering problem because of the limitations of the existing traditional clustering algorithms. In this chapter, the authors will give a brief review of various research papers that present the area of document or text clustering approaches with different metaheuristic algorithms.

INTRODUCTION

Exponential growth of text documents’ volumes is accelerated by a noticeable increase in digital libraries and repositories, social networking applications, company-wide intranets, digitized personal information such as blog articles and emails, etc. The effective usage of computers as well as Internet adds billions of electronic documents to the search area. This increase in both the volume and the variety of text documents requires advances in methodology to automatically understand, process, and summarize the data. Fast and high-quality cluster analysis plays an important role in helping users to effectively navigate, summarize and organize the large amount of information.

DOI: 10.4018/978-1-5225-0075-9.ch012
Cluster analysis has been the subject of many disciplines from past to present. Statisticians, engineers, psychologists, biologists, taxonomists, social scientists, mathematicians, computer scientists, medical researchers and others who are interested in real data have all contributed to cluster analysis. The classic definition of clustering is natural grouping of similar objects that exists in a set of patterns or data points while the similarities between objects in different groups are low. Different distance measures are used to quantify the similarity between two objects like euclidean distance, cosine distance, jaccard coefficient etc. Each measure has its own advantages and disadvantages that make it more or less suitable to a given domain or application area such as bioinformatics, document clustering or categorization. Moreover, choosing an appropriate similarity measure is also crucial for cluster analysis.

Clustering is a kind of unsupervised learning and should not be confused with classification, since unlike classification, no labelled documents are provided in clustering. However in some applications, cluster analysis is referred as unsupervised classification.

Document (or text) clustering, which is a subset of data clustering, is one of the most widely used topics in data mining researches. It includes concepts from the fields of information retrieval, automatic topic extraction, natural language processing, and machine learning. In document clustering, each document is considered as a vector in the term-space and according to documents’ similarity; each document is assigned to clusters. Although there are several ways to model a document in document clustering, using the frequency of each term as weight is the popular one.

Document-clustering algorithms are divided into a wide variety of different types in literature such as feature selection methods, distance based clustering algorithms, density based clustering algorithms, word and phrase based algorithms, probabilistic document clustering. On the other hand, hierarchical clustering algorithms and partitional clustering algorithms are the most commonly used ones. Different clustering algorithms have different tradeoffs in terms of effectiveness and efficiency.

Besides these traditional clustering algorithms, because of their robust, fast, and close approximate solution, metaheuristic algorithms are being used for document clustering in recent years.

In this study, first of all, a short view about metaheuristics is given. In addition to that, the authors briefly explain some topics related to the general document clustering procedure like document representation, similarity measures and evaluation of clustering solution. Afterwards, a review of various research papers that present the area of document or text clustering approaches with different metaheuristic algorithms will be summarized in a nutshell.

METAHEURISTICS

The solutions of real life problems can be numerous and sometimes an infinite number of solutions may be possible. In such a case, if the problem admits one solution, this will only actualize with a unique set of parameter values and traditional optimization approaches cannot be applied (Antoniou & Lu, 2007). On the other hand, the size of possible solutions that prevents an exhaustive search, the complexity and difficult constraints of the discussed problems, caused the approximate methods to be popular.

Metaheuristic algorithms, which are a class of approximate methods, have emerged in the 1980’s. The word “heuristic” has its origin in the old Greek word “heuriskein”, meaning the art of discovering new strategies (rules) to solve problems. The suffix “meta”, also a Greek word, means upper level