Effective Fuzzy Ontology Based Distributed Document Using Non-Dominated Ranked Genetic Algorithm

M. Thangamani, Kongu Engineering College, India
P. Thangaraj, Bannari Amman Institute of Technology, India

ABSTRACT

The increase in the number of documents has aggravated the difficulty of classifying those documents according to specific needs. Clustering analysis in a distributed environment is a thrust area in artificial intelligence and data mining. Its fundamental task is to utilize characters to compute the degree of related corresponding relationship between objects and to accomplish automatic classification without earlier knowledge. Document clustering utilizes clustering technique to gather the documents of high resemblance collectively by computing the documents resemblance. Recent studies have shown that ontologies are useful in improving the performance of document clustering. Ontology is concerned with the conceptualization of a domain into an individual identifiable format and machine-readable format containing entities, attributes, relationships, and axioms. By analyzing types of techniques for document clustering, a better clustering technique depending on Genetic Algorithm (GA) is determined. Non-Dominated Ranked Genetic Algorithm (NRGA) is used in this paper for clustering, which has the capability of providing a better classification result. The experiment is conducted in 20 newsgroups data set for evaluating the proposed technique. The result shows that the proposed approach is very effective in clustering the documents in the distributed environment.

Keywords: Clustering Analysis, Conceptual Clustering, Document Clustering, Genetic Algorithm, Ontology

1. INTRODUCTION

Clustering of document is not only an efficient way of organizing, summarizing, extracting and retrieving of information but also an important method in documentation. Initially, clustering is applied for enhancing the precision or in recalling information retrieval techniques. Clustering analysis is an important field of research intelligence and data mining. The fundamental concept of clustering analysis springs from characteristic to calculate the degree of similar relationship among objects and to attain automatic classification. Recently, clustering data in large data sets (Sharma et al., 2009) has gained a lot of attention in myriad research domains.

DOI: 10.4018/jiit.20111000102
Moreover, clustering technique is applicable in certain areas which involve browsing gathered data or in categorizing the outcome, provided by the search engine in terms of replies to the queries provided by the user. Document clustering can also be applied in producing the hierarchical grouping of document. In order to search and retrieve the information efficiently in Document Management Systems (DMSs), the metadata set should be created for the documents with enough details. One metadata per set is not enough for the document management systems. This is because the various document types need various attributes for distinguishing appropriately.

Though there are various document clustering techniques available in the literature but the existing clustering algorithms suffer from problems of practical applicability. The accuracy of the existing clustering approaches is a major concern. The time taken for active clustering of documents is more in large databases. So, a novel technique is needed for effective document clustering with very high accuracy but less complexity.

The discrete use of importance of using the ontology (Grootjen & Weide, 2005) increased in document clustering scheme as this technique seems to the accuracy of clustering. Whereas ontologies provide a position of legitimacy, research in this field leads to superior importance in the arising of different disputes featured in the modern digital situation. With the intention of providing efficacious solutions to the various drawbacks related to present search techniques, ontologies are widely implemented for creating viable document clustering techniques. An example of research concern in the field of ontology (Leukel & Sugumaran, 2009) relates to the Google which serves as search engine for semantic web documents, terms and data found on the internet. For retrieving information from the complex characteristic of the digital information available in digital libraries, ontologies can be successfully used for producing an efficient information retrieval system.

Ontologies (Murthy & Chowdhury, 1996) are introduced as modeling technology for structured metadata definition within the document clustering system. With the obtained metadata the clustering can be performed using the Genetic Algorithms (GAs) (Koller et al., 1997). GA is a search technique that is based on natural genetic and selection merging the idea of survival of the fittest with structured interchanges. Genetic algorithm (Sun et al., 2008) is a famous technique to deal with complex search problems by implementing an evolutionary stochastic search because genetic algorithm can be very effectively applied to various challenging optimization problems. The NP-hard nature of the clustering technique makes genetic algorithm a natural choice for solving it. A common objective function in these implementations is decreasing the square error.

Nowadays, document clustering in distributed environments have become a vibrant area of research. The huge amount of distributed high-dimensional data, such as digital libraries and web accessible text databases necessitate an infrastructure for effective similarity search in Peer-to-Peer (P2P) environments. P2P systems have become promising portals of solutions to deal with data management in case of high degree of distribution. Hence, the main challenge is to offer effective and scalable searching approach, in a highly distributed content, without essentially moving the actual content away from the information providers. Then the drawback of lack of global knowledge, i.e., each peer is conscious of only a small portion of the network topology and content has to be dealt with utmost care.

The principal aim of this research is the development of an improved distributed document clustering technique with very high classification accuracy. Moreover, the research also concentrates on the development of an effective document clustering technique with reduced classification time, less convergence time and less computational complexity. The main contribution of this paper is its usage of fuzzy ontology with the non-dominated ranked genetic algorithm.

This paper uses Ontology Generation using Fuzzy Logic (OGF) with non-dominated ranked genetic algorithm for effective document
Related Content

Formalization of User Preferences, Obligations and Rights
www.igi-global.com/chapter/formalization-user-preferences-obligations-rights/24175?camid=4v1a

Beyond Ambient Display: A Contextual Taxonomy of Alternative Information Display
www.igi-global.com/article/beyond-ambient-display/3877?camid=4v1a

Behaviour-Based Clustering of Neural Networks
www.igi-global.com/chapter/behaviour-based-clustering-neural-networks/10253?camid=4v1a
Iris Identification System: A New Perspective

www.igi-global.com/chapter/iris-identification-system/143564?camid=4v1a