ABSTRACT

In the current information explosion era, the amount of data stored in the form of text, image, video and audio is enormous and is expected to grow in future. Moreover, the technological breakthroughs in fields like e-libraries and e-publishing is increasing the magnitude and usage of digital documents. Thus, the tools that can be used to analyze and discover useful knowledge are in great demand. In particular, the usage of data mining techniques on text documents, also known as intelligent text analysis, text data mining or Knowledge-Discovery in Text (KDT) is becoming more popular. The reason behind this popularity is that most of the information (over 80%) is currently stored as text and therefore mining text information has a great commercial potential value. Document clustering helps users to effectively navigate, summarize and organize text documents into meaningful clusters, knowledge that helps to handle huge amount of text extraction. Pre-processing and feature selection are of extreme importance in document clustering. Modified Term Frequency/Inverse Document Frequency (TF/IDF) method is proposed for feature selection and these features are used by hybrid models that combine partition based and frequent item based algorithms with Ontology for efficient document clustering.
The KDT plays vital role in many applications like information extraction, concept/entity extraction, sentiment analysis, document summarization, entity-relation modeling, classification, and clustering. The existing techniques can be categorized as traditional and ontology-based clustering algorithms. Ontologies have been proposed to solve the problems that arise from high dimensionality and sparsity. Ontology-based algorithms allow the analyst to represent the complex structure of objects. The proposed work captures the advantage of both traditional and ontology-based algorithms for efficient text document clustering. The comparisons between traditional clustering, ontology-based clustering, and hybrid systems have not yet been probed early. This research applies the ontology framework to the document corpus; its different categorization method overcomes the challenges available in document clustering.

In the domain of document clustering, research is still at its peak. Ontology based methods allows the analyst to represent the complex structure of objects and to implement the relationship between categories and individual objects in hierarchical manner. There are many challenges involved in the existing traditional algorithms and the advantages provided by the ontology based algorithms have motivated this research work to enhance the existing techniques and combine them with ontology for clustering text documents.