Abstract

As the internet is exploding with huge volume of text documents, the need of grouping similar documents together for versatile applications have hold the attention of researchers in this area. Document clustering can facilitate the tasks of document organization and web browsing, search engine results, corpus summarization, documents classification, information retrieval and filtering. However several attempts have been made to develop efficient document clustering algorithms but most of the clustering methods suffer from challenges in dealing with problems of high dimensionality, scalability, accuracy and meaningful cluster labels. This paper intends to provide a brief summary over methods studied and current state of documents clustering research, including basic traditional methods as well as advanced fuzzy based, GA, PSO, HS oriented techniques etc. Also document representation model and its challenges, dimensionality reduction mechanisms, issues in document clustering, and cluster quality evaluation criteria are discussed.

References

summarization. UCAM-CL-TR-802 ISSN 1476-2986.

Index Terms

Computer Science

Artificial Intelligence

Keywords
document clustering hierarchical clustering partitioning clustering frequent item set vector space model