Abstract

Document clustering is an effective tool to manage information overload. By grouping similar documents together, we enable a human observer to quickly browse large document collections, make it possible to easily grasp the distinct topics and subtopics. In this Paper we survey the most important problems and techniques related to text information retrieval: document pre-processing and filtering, word sense disambiguation. Further we present text clustering using Modified FPF algorithm and comparison of our clustering algorithms against FPF, which is the most used algorithm in the text clustering context. Further we introduce the problem of cluster labeling: Cluster labeling is achieved by combining intra-clusterand
inter-cluster term extraction based on a variant of the information gain measure.

References

An Efficient Technique to Improve Snippet Clustering and Labeling using Modified FPF Algorithm

Index Terms

Computer Science

Algorithms

Keywords

Clustering  Document Clustering  Cluster Labeling  Information Retrieval