Abstract

Text document clustering plays an important role in providing intuitive navigation and browsing mechanisms by organizing large sets of documents into a small number of meaningful clusters. Clustering is a very powerful data mining technique for topic discovery from text documents. The partitional clustering algorithms, such as the family of K-means, are reported performing well on document clustering. They treat the clustering problem as an optimization process of grouping documents into k clusters so that a particular criterion function is minimized or maximized. The bag of words representation used for these clustering is often unsatisfactory as it ignores relationships between important terms that do not co-occur literally. In order to deal with the problem, we integrate core ontologies as background knowledge into the process of clustering text documents. This model combines phrases analysis as well as words analysis with the use of WordNet as background Knowledge and NLP to explore better ways of document representation for clustering. The Semantic based analysis assigns semantic weights to both document words and phrases. The new weights reflect the semantic relatedness between the documents terms and capture the semantic information in the documents to improve the web document clustering. The method adopted has been evaluated on different data sets with standard performance measures to develop meaningful clusters has been proved.
Text Document Clustering based on Semantics

References

- Tapas Kanungo, Nathan S. Netanyahu, Angela Y. Wu, « An efficient k-means clustering algorithm : Analysis and implantation »

Index Terms

Computer Science Data Mining

Keywords

Document Clustering K-means Semantic Weights Semantic Similarity Pos Tagging
Ontologies
Wordnet
Nlp
Similarity Measure