Abstract

Text mining is a modern and computational approach attempts to determine new, formerly unidentified information by pertaining techniques from normal language processing and data mining. Clustering, one of the conventional data mining techniques is an unsubstantiated learning pattern where clustering techniques attempt to recognize intrinsic groupings of the text documents, so that a set of clusters is formed in which clusters reveal high intra-cluster comparison and low inter-cluster similarity. Most current document clustering methods are based on the Vector Space Model (VSM), which is a widely used data representation for text classification and clustering. Moreover, weighting these features accurately also affects the result of the clustering algorithm substantially. The previous work described the conceptual text clustering to web documents, containing various mark up language formats associated with the documents (term extraction mode). In this work, we are going to present a Conceptual rule mining which is generated for the sentence meaning and related sentences in the document. Weights are appropriated for the sentences having higher contribution to the topic of the document. Conditional probability is evaluated for the sentence weights. Probability ratio is identified for the sentence similarity from which unique sentence meaning contributing to the document topic are listed. Experiments are conducted with the web documents extracted from the research repositories to evaluate the efficiency of the proposed efficient conceptual rule
mining on text clusters in web documents and compared with an existing Model for Concept Based Clustering and Classification in terms of Topic related rules, Weights of the influential sentence, Topic Sensitivity.

References

- Shady Shehata, Fakhri Karray, and Mohamed S. Kamel, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering", IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, October 2010.
- Yongzheng, Nur and Evangelos, "Narrative Text Classification for Automatic Key Phrase Extraction in Web Document Corpora", WIDM&apos;05, 51-57, Bremen Germany, 2005

Index Terms

Computer Science

Information Sciences
Keywords

Conceptual rule mining   text clustering   conditional probability   probability ratio