Text Document Clustering based on Phrase Similarity using Affinity Propagation

Abstract

Affinity propagation (AP) was recently introduced as an un-supervised learning algorithm for exemplar based clustering. In this paper novel text document clustering algorithm has been developed based on vector space model, phrases and affinity propagation clustering algorithm. Proposed algorithm can be called Phrase affinity clustering (PAC). PAC first finds the phrase by ukkonen suffix tree construction algorithm, second finds the vector space model using tf-idf weighting scheme of phrase. Third calculate the similarity matrix form VSD using cosine similarity. In Last affinity propagation algorithm generate the clusters. F-Measure, Purity and Entropy of Proposed algorithm is better than GAHC, ST-GAHC and ST-KNN on OHSUMED, RCV1 and News group data sets.

References

- RuiXu Donald C. Winch, "Clustering", IEEE Press 2009,pp 1-282
- Jain, A. and Dubes R. "Algorithms for Clustering Data", Englewood Cliffs,
Text Document Clustering based on Phrase Similarity using Affinity Propagation

- Renchu Guan, Xiaohua Shi, Maurizio Marchese, Chen Yang, and Yanchun Liang, “Text Clustering with Seeds Affinity Propagation”, IEEE Transaction on Knowledge and Data Engineering Vol. 23 No 4, 2011, pp 627-637.
Index Terms

Computer Science  
Artificial Intelligence

Keywords

text clustering  
affinity propagation  
unsupervised learning  
vector space model  
suffix tree  
tf-idf weighting scheme  
Purity  
Entropy-measure  
GAHC  
ST-GAHC  
ST-KNN