Abstract

Document clustering is used in information retrieval to organize a large collection of text documents into some meaningful clusters. k-means clustering algorithm of practical category, performs well on document clustering. k-means organizes a large collection of items into k clusters so that a criterion function is optimized. As it is sensitive to the initial values of cluster centroids, this paper proposes a graph-based method to calculate the appropriate initial cluster centroids. Document collection is represented as a graphical network in which a node represents a document and an edge represents the similarity between two documents. In order to calculate initial centroids, community structure present in graphical network is detected using edge deletion technique. Using community structure, centrality of each node is calculated. Centrality value of a node represents its candidacy of being a cluster centroid. Use of community structure assures that calculated centroids have sufficient number of topically related documents and centroids are well separated from each other. k-means with these initial centroids provides a significant improvement over simple k-means for text document clustering.
- Lerman K., Document Clustering in Reduced Dimension Vector Space., USC Information Science Institute, 4676 Admiralty Way, Marina del Rey, CA 90292.
- 20 Usenet newsgroups dataset, kdd.ics.uci.edu/databases
- The 4 universities data set, www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/

Index Terms

Computer Science

Artificial Intelligence
Keywords

Text mining  Document clustering  Cosine similarity  k-means