Abstract

Many users access the web seeking for information. They put their query or question in search engines that may return irrelevant pages or results compared to users' needs. This research paper proposes a model to remove outliers from the search results. The proposed model is based on association rules, modified Naïve Bayes algorithm and clustering techniques. The Naïve Bayes algorithm is modified to help removing outliers from the search results. The proposed model has been evaluated using the Sum of Squared Errors (SSE), silhouette coefficient and entropy evaluation measures against the standard k-medoids algorithm. Experimental results show that the proposed model outperforms the standard k-medoids clustering algorithm in removing the search outliers.

References


7. K. A. A. Nazeer, S. D. M. Kumar, and M. P. Sebastian, "Enhancing the k-means clustering algorithm by using a O(n logn) heuristic method for finding better initial centroids" , Second International Conference on Emerging Applications of Information Technology (EAIT), Kolkata, India, 2011.


11. H. Kim, X. Ren, Y. Sun, C. Wang, and J. Han, "Semantic frame-based document representation for comparable corpora", IEEE 13th International Conference on Data Mining (ICDM), Dallas, TX, USA, 2013.


18. J. Han, M. Kamber, and J. Pei, Cluster analysis: basic concepts and methods in Data mining concepts and techniques, Third Ed. New York, USA: Elsevier Inc.


21. J. Han, M. Kamber, and J. Pei, Classification: basic concepts in Data mining concepts and techniques. New York, USA: Elsevier Inc.


Index Terms

Computer Science

Information Sciences

Keywords

Information Retrieval (IR), Web mining, Association rules (AR), Classification, Clustering, Outlier detection.