Abstract

Now-a-days due to increase in the availability of computing facilities, large amount of data in electronic form is been generated. The data generated is to be analyzed in order to maximize the benefit of intelligent decision making. Text categorization is an important and extensively studied problem in machine learning. The basic phases in the text categorization include preprocessing features like removing stop words from documents and applying TF-IDF is used which results into increase efficiency and deletion of irrelevant data from huge dataset. Application of TF-IDF algorithm on dataset gives weight for each word which summarized by Weight matrix. Preprocessing reduces the size of dataset which ultimately improves the performance of search engine. After that, index is generated from dataset. Index contains term with its occurrence in file and also its location in file. This paper discusses the implication of efficient Information Retrieval system for text-based data using clustering approaches.
Developing an Expert IR System from Multidimensional Dataset

- Tian Xia, Yanmei Chai "An Improvement to TF-IDF: Term Distribution based Term Weight Algorithm"; JOURNAL OF SOFTWARE, VOL. 6, NO. 3, MARCH 2011
- Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth";
- Carlos Cobos, Henry Muñoz-Collazos, Richar Urbano-Muñoz, Martha Mendoza, Elizabeth Leónc, Enrique Herrera-Viedma "Clustering of web search results based on the cuckoo search algorithm and balanced bayesian information criterion"; ELSEVIER Publication, 2014 Elsevier Inc. All rights reserved, 21 May 2014.
- Rui Tang, Simon Fong, Xin-She Yang, Suash Deb, "Integrating nature-inspired optimization algorithms to k-means clustering"; 978-1-4673-2430-4/12/$31. 00 ©2012 IEEE.
- Carlos Cobos, Henry Muñoz-Collazos, Richar Urbano-Muñoz, Martha Mendoza, Elizabeth Leónc, Enrique Herrera-Viedma "Clustering Of Web Search Results Based On The Cuckoo Search Algorithm And Balanced Bayesian Information Criterion"; ELSEVIER Publication, 2014 Elsevier Inc. All rights reserved, 21 May 2014.
- http://www. ardentertat. com/2011/05/30/how-to-implement-a-search-engine-part-1-create-index/
Developing an Expert IR System from Multidimensional Dataset

Index Terms

Computer Science  Multimedia

Keywords

Information retrieval  stop words  TF IDF  text based clustering  fitness functions