Similarity based Retrieval of Documents using Jaccard Coefficient Similarity Methods

Nan Ei Khaing, Sabai Phyu
University of Computer Studies, Yangon
nangeikhaing20th@*gmail.com,*
sabaiphyu@ucsy.edu.mm

Abstract

Particularly, information retrieval results as documents are typically too extensive. Consequently, a similarity measurement between keywords and index term is essentially performed to facilitate searchers in accessing the required results. Thus, this paper proposed the similarity measurement method between words by deploying Jaccard Coefficient. After documents are collected, some pre-processing tasks such as stop-word removal, stemming, and handling of digits, hyphens, punctuation, and cases of letters are usually performed. Almost 200 documents (with pdf formats) from UCSYare conducted in this system. With the help of this system, the users can search effectively the most related documents with their preference information without boring. Furthermore, the performance of the proposed similarity measurement method: Jaccard similarity coefficient was accomplished by employing precision, recall, and F- measure.