Two Steps Break-Cull Model for Automatic Indexing of Persian Texts

Mohammad Tavakolizadeh-Ravari
Assistant Professor of KIS, Yazd University
tavakoli@yazd.ac.ir
Received: 26th October 2013; Accepted: 10th June 2014

Abstract

Purpose: Each language has its own problems. This leads to consider appropriate models for automatic indexing of every language. These models should concern the exhaustificity and specificity of indexing. This paper aims at introduction and evaluation of a model which is suited for Persian automatic indexing. This model suggests to break the text into the particles of candidate terms and to cull the most appropriate ones through a special method of term weighting.

Methodology: The introduction method of the automatic indexing model is performed through showing the steps and the possible problems for running them. Evaluation is based on the inclusion index. This index is used for determination the inter-indexer consistency. Therefore, the consistency of resulted index terms (from this model) and author keywords is determined.

Findings: Findings show that 90% of articles' most weighted terms are similar to their first author keywords. The overall consistency between the results of running the model and author keywords is 76%. Compared with the prior works, the performance of the model is acceptable.

Originality/Value: The initial value of this paper is concerning the automatic indexing with regard of Persian language problems. The model is well suited for using regular expression language which is supported by many programming languages. This diminishes the need to create database tables for text manipulation and processing. In addition, the model solves the problem of upper threshold for determination of final terms. Another algorithm makes it possible to determine the lower one. Finally, the number of culled terms does not depend on the text length. This guaranties the exhaustificity and specificity of indexing.

Keywords: Automatic Indexing; Persian Language; Break-Cull Model.