APLIKASI PENCARIAN FRASE DARI SEKUMPULAN DOKUMEN

Susana Limanto

[1)] Teknik Informatika Universitas Surabaya
Jl. Raya Kalirungkut Surabaya - 60292
email : susana@ubaya.ac.id

ABSTRACT

It is difficult to find a specific document from documents that are not well organized. One solution is to provide a label to each set of documents that best describes the content of the document. Phrases that are most often used in those documents can be used as a label.

Corephrase is a method that can be used to extracting phrases from documents. After phrases were extracted, their scores is calculated. Score from every phrase is computed based on document frequency, average phrase frequency, average weight, and average phrase depth. Finally, all of the scores are rank by sorting their scores. The top rank is the representative label.

In this research, an application progam is built to extracting phrases from documents by using the corephrase method. Documents and stopword list is required as application's input. This aplication will generate top ten phrases as candidate label.

In order to know correctness of the corephrase method on the application program, a testing was made on scientific documents and then compared our manual calculations with the results given by the application program. After several improvements, experimental results show that the application is running correctly. Experimental result also shows that the stemming process does not always work well. It means that porter stemming does not always give perfect stems.

Key words

Phrase, Metode Core phrase, Porter Stemmer