PSEUDO-RELEVANCE FEEDBACK ON RETRIEVAL USING DOCUMENT SEGMENTATION

Elenur Dwi Anbiana, Julio Adisantoso

Department of Computer Science

Faculty of Mathematics and Natural Sciences

Bogor Agricultural University

julioipb@gmail.com

2009

Abstract

Web is the largest information source in the world, but the storage and effective information retrieval on the web is still a problem in this day. Pseudo-relevance feedback is automatically local analysis technique (relevance feedback technique without explicit input user). This technique takes the top n-ranked documents as the relevant documents and takes the top x-ranked terms from relevant documents. Basically, a document consist of any topics, so in the research, relevant documents substituted by a segment which represents a topic in the relevant document. The segment is taken by XML document tag, are {TITLE}, {AUTHOR} and {P} tag, since the text from those tags are usually represent a document. The visual block extraction phase (first step in VIsion based Page Segmentation algorithm), used in segmented the document, so the title of the research is Pseudo-Relevance Feedback on Retrieval Using Document Segmentation. This research done in six phases, there are retrieval initialization, document segmentation, segments selection, terms selection, final retrieval and retrieval output evaluation. The result of system performance is good, that is 0.5214. The test results show that the performance of the PRF based segmentation of documents and without PRF retrieval was not found significant differences. It is because of the taken documents and segments for expansion terms selection are not relevant, expanded queries are not exact representing the segment, the member of documents that are used in retrieval is relatively small, that are 1000 documents.