Abstract

A search engine must be able to precise a relevant result that the user demand. Search engine is a system that used information retrieval concept. In information retrieval, there are two types of documents, namely the free text (unstructured document) and fielded text (structured document). HTML document is one kind of the fielded text. Searching in HTML documents must consider a factor of importance of each part of the document. Those factors, hereinafter referred as static rank, can be distinguished based on tags or markup in HTML documents, such as title, text, inlinks, obj, type, etc.

BM25F method is one kind of method that applying the IR score and static rank. BM25F method is implemented in the scope of the document weighting, which having a certain calculation of the field (tag) that is affected by certain boost factors. The performances of this method are based on suitability BM25F documents with keywords (queries), hereinafter referred as value relevance.

By applying BM25F method it was discovered that, with changes in the large number of documents (N) and the number of relevant document to the proportion remains in use, the value of performance based on precision and recall of information retrieval system using BM25F method produces a good performance, which is likely to approach a stable state.

Based on results of testing on boost factor scenario, it can be concluded that the results of system performance (precision, recall and IAP / interpolated average precision) is best obtained when using the default scenario ({title = 4}; {H1 = 3}; {anchor = 2};} {div = 2; {text = 1};). This is caused by the given proportion off boost factor value equal to the default scenario, in order of importance of each field in the document based on assumptions. However, there is another factor that must be considered. This is the diversity of the field (tag) existence which is tested in the scenario in each of documents.

Keywords : information retrieval, information retrieval system, BM25F