ABSTRACT

Information retrieval (IR) is a search for information that is usually in a text document. In this study discussed the information retrieval of the Indonesian translation of the Quran consisting of 6236 verses and the Hadith of Sahih Bukhori consisting of 7008 Hadith. The technique used is a similarity calculation with Cosine Similarity and use weighting TF-IDF Vector Space Model (VSM). The main purpose of information retrieval is search and display relevant documents according to queries. Problems with Information retrieval that is often found is that documents that are considered to be less relevant are ranked in the top search results. This research is testing for improve the relevance and precision of the results of information retrieval searches. Methodology in this study consisted of stages namely the formation of the Corps of the Qur'an and Hadith in Indonesian translation, formation of Corpus Synonyms (thesaurus) and the formation of themes from the verses of the Qur'an and Hadith. The algorithm used is the development of Cosine Similarity at Term Frequency and Inversion Document Frequency TF-IDF Vector Space Model (VSM). The development carried out is to add 2 criteria as Cosine Similarity calculation component, namely the principle "and" the most without duplicates and words that in line / sequentially according to the keyword (query) to be prioritized to be ranked on the calculation of document similarity values. Sequential words are a priority can also solve idiom problems (expressions) because idioms consist of 2 words or more that usually has 1 meaning (word). "By adding these two criteria will further add relevance and precision in a bag of word (Token) method. Testing is done by testing the Al Quran verse search in the application of information retrieval and comparing the results of the search application with expert opinion Al Quran and Hadith. The first test is testing the search results with corpus synonyms (thesaurus). The results of this test show that in the presence the results of the corpus synonym (thesaurus) are broader with very additions significant; testing by entering keywords using 1 word 2 words and 3 words or more (a sentence). The average test results in a recall reaching 100% and 85% precision with F-Measures 92%. Trials theme classification in Al Quran recall value of 75%, 53% precision and F-Measures 61%. This research has proven that the development of the cosine similarity algorithm in VSM TF-IDF and uses corpus synonyms (thesaurus), increases the level relevance and precision, because it significantly expands the search results and reduces irrelevant documents.

Keywords: Al Quran, Hadith, Corpus, Information Retrieval, UGIrfanCosim, TF-IDF, VSM, Cosine Similarity, Thesaurus