ABSTRACT

Most search engines this days generally displaying document of search results based on document order (document ranking) without grouping or categorize documents based on similarity. With a large number of documents, this will make a negative impact for users, it takes a relatively long time to sort out the documents that user needs. To facilitate the user in finding information on a large number of documents, one of the solutions is to classify document of search results according to the keywords entered by the user. With the document search results are grouped, the user does not need to open up too much pages because the document of search results have been grouped based on documents similarity.

One partitional algorithm that able to classify unlabeled documents is Expectation-Maximization, this algorithm used to find the value of Maximum Likelihood estimation of parameters in a probabilistic model [2]. The characteristics of this algorithm is able to classify documents that have not been labeled or unlabeled documents and also the results of the classification will always convergence. From the experimental results it was concluded that the EM algorithm can classify documents of search results, it can help users to search for documents they expected. The highest accuracy reaches 70% and the lowest 32.58%. The addition of stemming algorithms Arifin Setiono EM algorithm can improve performance up to 10%.

Keyword: Clustering, Expectation-Maximization, Unsupervised, Stemming.