PENGELOMPOKAN ARTIKEL BERBAHASA JAWA DENGAN *HIERARCHICALK* MEANS*CLUSTERING*

ABSTRAK

Artikel memiliki berbagai jenis topik, sebagai contoh: berita ekonomi, kesehatan, dan sebagainya. Berdasarkan pada jenis artikel di atas ternyata dapat digali informasi yang dapat dimanfaatkan (*knowledge discovery). Knowledge discovery*pada data teks dapat dilakukan dengan proses awal berupa *information retrieval.* Proses dari *information retrieval* bertujuan untuk menemukan ciri dari dokumen, untuk selanjutnya dilakukan analisis keterhubungan antar dokumen dengan menggunakan metode pengelompokan. Sebelum dikelompokkan, data dokumen dari media cetak harus diubah ke bentuk *text file*. Selanjutnya masuk tahap *information retrieval*untuk memperoleh ciri dari suatu dokumen. Proses yang dilakukan adalah *tokenizing*, *stop word*, *stemming*, dan *weighting.* Berdasarkan proses *information retrieval* yang telah dilakukan, data dikelomopokan menggunakan *Hierarchical K Means.* Metode *Hierarchical K Means* terdiri dari dua buah algoritma utama, yaitu *K Means* dan *agglomerative hierarchical clustering* (AHC) khususnya teknik *single linkage. Single linkage* dilakukan mencari *centroid* yang paling baik. Proses selanjutnya dilakukan *K Means* dengan menggunakan *centroid* hasil *single linkage*, guna menghasilkan *cluster* terbaik. Setiap hasil *cluster* dievaluasi dengan metode evaluasi internal,metode yang digunakan adalah *sum of square error* (SSE). *Cluster* yang memiliki error minimum diuji kembali dengan evaluasi eksternal, yaitu dengan menggunakan (*confusiion matrix*). Berdasarkan percobaan pengelompokan yang dilakukan didapatkan pembentukan tiga *cluster,* yang memiliki error *cluster* minimum 19,84822 (evaluasi internal) dan memiliki akurasi maksimum 80% (evaluasi eksternal). Pembentukan tiga kelompok ini juga sesuai dengan tujuan yang ingin dicapai dalam tulisan ini, yaitu untuk mendapatkan pengelompokan dari artikel dan dapat membantu untuk mengetahui jenis topik artikel.

JAVANESE LANGUAGEARTICLESCLUSTERING USING HIERARCHICAL KMEANS

ABSTRACT

There are many kinds of topic article—economy, health, politic, etc. Within those articles, there is useful information that can be found (knowledge discovery). Knowledge discovery on the text data could be initiated by the initial process called information retrieval. The information retrieval process aimed to collect the characteristic of a document in order to analyze the connection between documents by using clustering method. Before conducting the clustering process, document'sdata from printed media should be converted into text file. The next step is information retrieval. In this step, the information retrieval collected the characteristic of a document by using tokenizing, stop word, stemming, and weighting. Documents data clustered by using Hierarchical K Means method based on information retrieval. This method consisted of two main algorithms, which are K Means and agglomerative hierarchical clustering (AHC) with single linkage technic. Single linkage would collect the best centroid. In the next process, K Means was initiated using best centroid from AHC to produce best cluster. Every cluster produced would be evaluated by internal evaluation method. The internal evaluation method is sum of square error (SSE). Clusters with minimum error would be retested by external evaluation method using confusion matrix. There are three outcome of clusters based on the clustering trial, which have minimum error 19,84882 (internal evaluation) and maximum accuracy 80% (external evaluation). The forming of these three clusters was corresponded with this paper's objectives, which are tocluster the article and to find out the type of the article topic.