ABSTRAK

Bahasa Jawa merupakan salah satu bahasa daerah di Indonesia yang sangat sering digunakan. Banyak artikel Bahasa Jawa yang dapat kita jumpai setiap hari dalam bentuk dokumen digital. Untuk mempermudah seseorang dalam penemuan informasi dalam artikel Bahasa Jawa yang dicari dapat dilakukan dengan menggunakan klasifikasi dokumen. Penelitian ini bertujuan untuk membuat suatu aplikasi yang mampu mengklasifikasikan artikel bahasa Jawa menggunakan sistem pemerolehan informasi dan dikombinasikan dengan algoritma *K-Nearest Neighbor*.

Penelitian ini membagi dokumen ke dalam empat kategori yaitu politik, ekonomi, kesehatan, dan pendidikan. Proses klasifikasi dokumen diawali dengan membaca dokumen, tokenisasi, *stopword*, *stemming*, *text frequency*. Sistem ini menggunakan vektor ciri TF-IDF (*term frequency/ Inverse document frequency)*. *Term frequency* adalah jumlah kemunculan suatu kata dalam sebuah dokumen, sedangkan *inverse document frequency* adalah *inverse* dari banyaknya dokumen dimana suatu term tersebut muncul. Setelah menghitung TF-IDF dilakukan perhitungan *Cosine Similarity*. *Cosine Similarity* merupakan algoritma yang digunakan untuk menghitung kemiripan antara dokumen baru dan dokumen pelatihan. Untuk melakukan klasifikasi dokumen digunakan algoritma *K-Nearest Neighbor*. Metode *K-Nearest Neighbor* mengklasifikasikan dokumen dengan menggunakan hasil dari perhitungan TF-IDF yang digunakan untuk menghitung kedekatan antar dokumen (*cosine similarity*)

Pada penelitian ini dilakukan pengujian yaitu dengan *cross validation* kemudian dilakukan uji presisi. Data yang digunakan sebanyak 40 dokumen. Tingkat akurasi untuk 3 *fold* k = 4 mencapai 95% dan k = 8 mencapai 92%, untuk 5 *fold* k = 4 mencapai 92% dan k = 8 mencapai 94%.
.
Kata kunci : klasifikasi dokumen bahasa Jawa, *K-Nearest Neighbor*, *K-NN*, pemerolehan informasi

ABSTRACT

Javanese language is one of local / traditional languages in Indonesia which is always used. There are many Javanese language articles that always can be found in digital document form. Clasification document can be used to find information in Javanese . The purpose of this research is to create an aplication which is able to clasify Javanese language article by using the combination of the information retrieval system and K-Nearest neighbor algorithm.

This research divided the documents into 4 categories which consist of : politic, economy, health and education. The process of clasification begins with reading the document, tokenizing, stopword, stemming, text frequency . The system uses a feature vector is TF-IDF (term frequency/inverse document frequency). Term frequency is the sum of a word's frequency in one term, meanwhile, inverse document is the frequency of documents in one term. Cosine similarity will calculate after calculating TF-IDF . Cosine similarity is the algorithm which is used to calculate similarity between the new document and the exercise document. K-Nearest Neighbour algorithm is using to clasify the document. K-Nearest Neighbor methode clasified the document by using the equal of calculating TF-IDF is used to compute the proximity between documents (cosine similarity).

This research also tested by cross validation then presision test. Using 40 data of documents. Accurancy for 3 fold k = 4 reaches 95 % and k = 8 reaches 92%, for 5 fold k =4 reaches 94 % and k = 8 reaches 94% .

Keywords : Javanese languange classification, K-Nearest Neighbor, K-NN, Information Retrieval