Pengelompokan dokumen bahasa indonesia dengan teknik reduksi dimensi nonnegative matrix factorization dan random projection

Suryanto Ang, author

Abstrak

Pengelompokan dokumen atau document clustering telah menjadi suatu teknik yang berguna dalam pengorganisasian sekumpulan dokumen. Dengan teknik ini, komputer bisa secara otomatis mengelompokkan sekumpulan dokumen ke dalam kluster-kluster yang cocok yang merepresentasikan data yang ada. Dengan demikian, proses pencarian informasi bisa dilakukan dengan lebih efisien. Telah banyak metode yang dikembangkan untuk mendukung pengelompokan dokumen. Dua diantara metode-metode tersebut adalah Nonnegative Matrix Factorization (NMF) dan Random Projection (RP). Pada penelitian ini, proses pengelompokan dokumen dilakukan dengan metode reduksi dimensi NMF dan RP pada dokumen berbahasa Indonesia. Untuk metode RP, diperlukan tahap tambahan untuk dapat mengelompokkan dokumen. Metode yang digunakan pada tahap ini adalah K-Means. Data yang digunakan pada percobaan adalah artikel media massa. Percobaan dilakukan dengan variasi pada variabel percobaan seperti jumlah kluster, jumlah data, jenis data, dan informasi fitur.

Dari percobaan yang telah dilakukan, terlihat bahwa teknik NMF dan RP dapat diterapkan dalam aplikasi pengelompokan dokumen bahasa Indonesia. Akurasi pengelompokan bisa mencapai 97%. Dari percobaan terlihat juga bahwa teknik NMF menghasilkan akurasi yang lebih tinggi daripada RP dengan kisaran perbedaan sekitar 2%. Ukuran dan jumlah kluster juga mempengaruhi akurasi. Ukuran kluster yang semakin besar menyebabkan peningkatan akurasi sedangkan jumlah kluster yang semakin banyak menyebabkan penurunan akurasi. Dengan ukuran kluster 296 dan jumlah kluster 2 misalnya, akurasi mencapai 96%. Disamping itu, informasi fitur berupa presence merupakan yang paling cocok digunakan karena menghasilkan akurasi yang paling tinggi, juga mencapai 97%. Jumlah fitur yang lebih banyak dan tidak mengandung stopwords juga memberikan akurasi yang lebih tinggi.
<hr>
Document clustering has been a beneficial technique in organizing documents. With good document clustering technique, computer can automatically group collection of documents into meaningful clusters. The information retrieval process thus can be done eficiently. There have been lots of methods developed in supporting document clustering process. Two of them are Nonnegative Matrix Factorization (NMF) and Random Projection (RP). In this research, document clustering process is conducted on Indonesian documents using both NMF and RP dimensional reduction method. For RP, additional clustering process is required. For this purpose, K-Means is used. Documents used are mass media articles. Experiments are conducted with variation of experiment variables including number of cluster, number of data, types of data, feature, etc.

From the experiments conducted, it can be concluded that NMF and RP technique can be used in document clustering application for Indonesian documents. The accuracy reaches 97%. Experiments also show that

NMF yields better accuracy than RP with difference range about 2%. Cluster size and cluster number also influence the accuracy. The bigger the cluster size, the higher the accuracy while the more the cluster number, the lower the accuracy. For example, with cluster size 296 and cluster number 2, the accuracy reaches 96%. Despitefully, using presence as feature is the most appropriate one because it results in the highest accuracy among others, also reaches 97%. In addition, the more the features used and excluding the stopwords, the higher the accuracy will be.