Web news documents clustering in indonesian language using singular value decomposition-principal component analysis and ant algorithms

---

Abstrak

Ant-based document clustering is a cluster method of measuring text documents similarity based on the shortest path between nodes (trial phase) and determines the optimal clusters of sequence do-cument similarity (dividing phase). The processing time of trial phase Ant algorithms to make docu-ment vectors is very long because of high dimensional Document-Term Matrix (DTM). In this paper, we proposed a document clustering method for optimizing dimension reduction using Singular Value Decomposition-Principal Component Analysis (SVDPCA) and Ant algorithms. SVDPCA reduces size of the DTM dimensions by converting freq-term of conventional DTM to score-pc of Document-PC Matrix (DPCM). Ant algorithms creates documents clustering using the vector space model based on the dimension reduction result of DPCM. The experimental results on 506 news documents in Indo-nesian language demonstrated that the proposed method worked well to optimize dimension reduction up to 99.7%. We could speed up execution time efficiently of the trial phase and maintain the best F-measure achieved from experiments was 0.88 (88%).

Klasterisasi dokumen berbasis algoritma semut merupakan metode klaster yang mengukur kemiripan dokumen teks berdasarkan pencarian rute terpendek antar node (trial phase) dan menentukan sejumlah klaster yang optimal dari urutan kemiripan dokumen (dividing phase). Waktu proses trial phase algoritma semut dalam mengolah vektor dokumen tergolong lama sebagai akibat tingginya dimensi, karena adanya masalah sparseness pada matriks Document-Term Matrix (DTM). Oleh karena itu, penelitian ini mengusulkan sebuah metode klasterisasi dokumen yang mengoptimalkan reduksi dimensi menggunakan Singular Value Decomposition-Principal Component Analysis (SVDPCA) dan Algoritma Semut. SVDPCA mereduksi ukuran dimensi DTM dengan mengkonversi bentuk freq-term DTM konvensional ke dalam bentuk score-pc Document-PC Matrix (DPCM). Kemudian, Algoritma Semut melakukan klasterisasi dokumen menggunakan vector space model yang dibangun berdasarkan DPCM hasil reduksi dimensi. Hasil uji coba dari 506 dokumen berita berbahasa Indonesia membuk-tikan bahwa metode yang diusulkan bekerja dengan baik untuk mengoptimalkan reduksi dimensi hingga 99,7%, sehingga secara efisien mampu mempercepat waktu eksekusi trial phase algoritma se-mut namun tetap mempertahankan akurasi F-measure mencapai 0,88 (88%).