INTISARI

CLUSTERING TEKS DENGAN ALGORITMA LINGO MENGGUNAKAN LATENT SEMANTIC INDEXING PADA CLUSTER CONTENT DISCOVERY

Muh. Bagus Tesa Karuniawan
12/334603/PA/14836

Hasil pencarian dokumen teks pada umumnya disajikan dalam bentuk daftar panjang yang diurutkan berdasarkan relevansinya terhadap *query* yang diberikan. Hasil pencarian tersebut tentu saja mengabaikan kedekatan antar dokumen itu sendiri. Kemiripan antar dokumen dapat disajikan melalui algoritma pengelompokan (*clustering*) yang mengelompokkan dokumen berdasarkan topik-topik tertentu.

Penelitian ini mengeksplorasi Algoritma LINGO untuk mengelompokkan dokumen teks dengan menggunakan *latent semantic indexing*. Algoritma ini terdiri atas 5 tahap. Tahap pertama adalah *preprocessing* yang melibatkan identifikasi bahasa, tokenisasi, serta *stemming*. Tahap kedua merupakan *feature extraction* untuk menentukan kandidat label untuk *cluster* yang akan dibentuk. Tahap selanjutnya adalah *cluster label induction* dimana kandidat label diseleksi menggunakan reduksi dimensi. Setelah kandidat label diseleksi, dilanjutkan dengan tahap *cluster content discovery* untuk menentukan keanggotaan kelompok masing-masing dokumen. Pada tahap *cluster content discovery* dalam penelitian ini digunakan *latent semantic indexing*. Penggunaan LSI pada tahap *cluster content discovery* memungkinkan pengelompokan dokumen tidak hanya didasarkan pada *lexical matching* dari label kelompok yang ada. Tahap kelima merupakan finalisasi kelompok yang terbentuk dalam bentuk perhitungan skor kelompok yang terbentuk.

Pengujian dengan mengelompokkan 5 hingga 135 dokumen yang diambil secara acak dari 300 dokumen tersedia menunjukkan bahwa penggunaan *latent semantic indexing* memberikan keberhasilan terbatas. Jumlah *cluster* yang hanya memiliki satu anggota saja semakin sedikit. Implementasi Algoritma LINGO tanpa LSI memberikan rata-rata *single cluster* sebesar 14,01. Sedangkan penggunaan LSI pada tahap *cluster content discovery* menghasilkan rata-rata jumlah *single cluster* sebesar 2,59. Meski demikian, terdapat lebih banyak kelompok yang saling tumpang tindih mengakibatkan nilai evaluasi *dunn index*, *partition coefficient*, serta *partition entropy* menjadi lebih buruk. Pengujian tanpa *latent semantic indexing* memberikan *dunn index* dengan rata-rata 0,37 sedangkan penggunaan LSI 0,17. Rata-rata nilai *partition coefficient* tanpa *latent semantic indexing* adalah 0,71 dan 0,49 ketika menggunakan *latent semantic indexing*. Evaluasi *partition entropy* memberikan rata-rata 0,10 untuk implementasi tanpa LSI dan 0,18 untuk implementasi dengan *latent semantic indexing*.

Kata Kunci: Clustering, LINGO, Latent Semantic Indexing

ABSTRACT

TEXT CLUSTERING WITH LINGO ALGORITHM USING LATENT SEMANTIC INDEXING ON CLUSTER CONTENT DISCOVERY

Muh. Bagus Tesa Karuniawan
12/334603/PA/14836

Text document search results mostly presented as a long list of documents in descending order of relevance. The way search result being presented disregard the similarity of the documents itself. Clustering algorithm is capable to present text documents relations in topical manner.

This research explores LINGO algorithm to cluster text document with latent semantic indexing. This algorithm consist of 5 major steps. Preprocessing stage covers language detection, tokenization, and stemming. The second stage is the feature extraction to determine the candidates for cluster labels. The next stage is the cluster label induction where candidates are selected through dimension reduction over the term-document matrix. Once the candidate labels are selected, cluster content discovery phase determines the membership of each documents. At this stage, latent semantic indexing is being used. Hence the discovery of documents membership is not only based on lexical matching to given labels. Finalization stage calculates the final score for given clusters.

The implementation tested with 5 to 135 documents taken randomly from 300 available documents. The test shows that latent semantic indexing offers limited success on addressing several issues in LINGO. The number of cluster that only consist of single document greatly reduced. LINGO implementation without LSI gives an average number of single cluster of 14.01. Latent semantic indexing reduces it into 2.59 on average. However, the number of overlapping clusters worsen cluster evaluation index. Tests without latent semantic indexing gives dunn index score 0.37 on average, but falls into 0.17 with latent semantic indexing employed. On average, partition coefficient without latent semantic indexing score around 0.71. LINGO that uses latent semantic indexing only able to score 0.49 on average for partition coefficient. Partition entropy also scores worse with 0.18 with LSI while it can score 0.10 on average without LSI.

Keywords: Clustering, LINGO, Latent Semantic Indexing