ANALISIS DAN IMPLEMENTASI PENGELOMPOKAN HASIL PENCARIAN MENGGUNAKAN ALGORITMA LINGO

Moh. Agus Sulistiono[1], Yanuar Firdaus A.w.[2], Retno Novi Dayawati[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Pencarian dokumen yang ada pada saat sekarang ini adalah dengan menampilkan hasil pencarian terurut berdasarkan peringkat kecocokan (document ranking). Hasil yang ditampilkan terkadang tidak sesuai (relevan) dengan yang diinginkan oleh pengguna.

Salah satu cara untuk mengelompokkan dokumen adalah dengan clustering. Pada Tugas Akhir ini dilakukan pengelompokkan dokumen berbahasa Indonesia dari dokumen koleksi dengan algoritma LINGO. LINGO merupakan algoritma clustering yang lebih mengedepankan kualitas penamaan label pada klaster.

Setelah implementasi, algoritma ini bisa membentuk klaster dengan dokumen-dokumen di dalamnya sesuai dengan labelnya, hal ini dikarenakan setiap dokumen dialokasikan ke masing-masing klaster berdasarkan tingkat kemiripannya dengan label yang terbentuk. Dalam menentukan label untuk penamaan klaster, algoritma ini memeriksa kemunculan term atau complete phrase dalam dokumen. Maka dari itu, algoritma ini sangat efektif jika dokumen-dokumen yang diproses banyak mengulang topik inti, sebaliknya akan kurang efektif jika topik inti dari dokumen diinterpretasikan dengan berbagai istilah yang beragam.

Dalam pengujian label klaster dengan metode precision dan recall pada metode pembobotan Term Frequency (TF) dan Term Frequency – Inverse Document Frequency (TF-IDF) didapatkan hasil yang bagus untuk keduanya

Kata Kunci : klaster, clustering, LINGO, complete phrase, precision, recall

Abstract

Nowadays, when searching documents, the search result will sort retrived documents based on their rank. The results sometimes irrelevant and different from user's expectation. One alternative to improve the search results is to clusterize it.

Documents in this final project will be document collection in Indonesian language and be clustered using LINGO algorithm. LINGO is clustering algorithm which ensure that both contents and description (labels) of the resulting groups are meaningful to the users. After implementation, this algorithm produce clusters that contains relevant documents to cluster label due to for each document is allocated to the cluster based on it's similarity to the label cluster. To determine the labels to describing cluster, this algorithm will check the occurrence of the term and complete phrase in the documents. So the algorithm will become effective if processed documents contains recurrent topic terms, otherways it will become uneffective if topic terms or phrases in the documents are interpreted in many different terms.

Due cluster label testing process using precision and recall on Term Frequency (TF) weighting and Term Frequency – Inverse Document Frequency (TF-IDF) generate good result for the both weighting methods

Keywords : cluster, clustering, LINGO, complete phrase, precision, recall

1. Pendahuluan

1.1 Latar belakang

Sedemikian pesatnya pertambahan jumlah dokumen beserta keanekaragamannya, menyebabkan masalah baru pada saat pencarian dokumen. Salah satu kesulitan tersebut yaitu mendapatkan hasil pencarian yang relevan.

Pada kebanyakan mesin pencarian saat ini, respon dari *query* pengguna mengembalikan hasil pencarian dengan menampilkan sebagian dari dokumen (*snippets*). Jika *query* terlalu umum, maka sangat sulit bagi user untuk mengidentifikasi dokumen mana yang sesuai. Pengguna diharuskan untuk melihat satu-persatu detil hasil pencarian dokumen untuk mengetahui dokumen mana yang relevan bagi user. Dan juga, keterhubungan antar dokumen pada hasil pencarian tidak disediakan.

Salah satu alternatif yang dapat menyelesaikan masalah diatas yaitu pengelompokan hasil pencarian secara otomatis ke dalam kelompok-kelompok tematik (*cluster*). Hal ini dapat membantu pengguna dalam mengidentifikasi dokumen hasil pencarian secara spesifik.

Penamaan klaster yang mudah dimengerti (*readable*) merupakan hal yang penting dalam menentukan kualitas dari *clustering*. Hal tersebut dapat membantu pengguna untuk mengidentifikasi kelompok dokumen yang dicari. LINGO (*Label INduction Grouping algOrithm*) merupakan algoritma untuk pengelompokan hasil pencarian yang mengedepankan kualitas penamaan klaster.

Dalam Tugas Akhir ini diterapkan suatu algoritma LINGO dalam aplikasi pengelompokan data. Alasan penggunaan algoritma ini adalah karena kemampuannya dalam menentukan kandidat label tidak hanya didapat dari kemunculan kata terbanyak dalam suatu klaster, namun bisa mendapatkan suatu frase atau gabungan kata dengan menggunakan metode *suffix array* sehingga penamaan suatu klaster lebih bisa dimengerti *user*. Dalam implementasinya, LINGO menggabungkan metode-metode antara lain *Vector Space Model* yang memodelkan dokumen sebagai sebuah vektor dengan bobot sesuai *term*. Juga dipakai *Latent Semantic Indexing* yang dikenal dengan kemampuannya dalam mendapatkan inti konsep dari dokumen dengan membangun matriks asosiasi antara *term-term* yang muncul pada konteks yang sama. Pada proses *clustering pada* umumnya, anggota klaster dicari terlebih dahulu kemudian *label* ditentukan. Namun pada LINGO proses tersebut dibalik, yaitu menentukan *label* dan jumlah klaster terlebih dahulu baru kemudian menentukan anggota-anggota pada *label* klaster yang telah ditentukan. Hal ini juga yang membuat pemilihan *label* klaster bisa mewakili keseluruhan topik dokumen.

1.2 Perumusan masalah

Masalah yang diteliti berdasarkan latar belakang diatas adalah sebagai berikut:

1. Bagaimana menerapkan algoritma LINGO pada aplikasi pengelompok hasil pencarian.
2. Menganalisis bagaimana performansi algoritma LINGO terhadap kesesuaian penamaan suatu klaster dengan isi dokumen-dokumen di dalam *klaster* tersebut.

Dalam penelitian Tugas Akhir ini, objek penelitian dibatasi dengan ruang lingkup sebagai berikut:

1. Dokumen yang digunakan hanya dokumen berbahasa Indonesia.

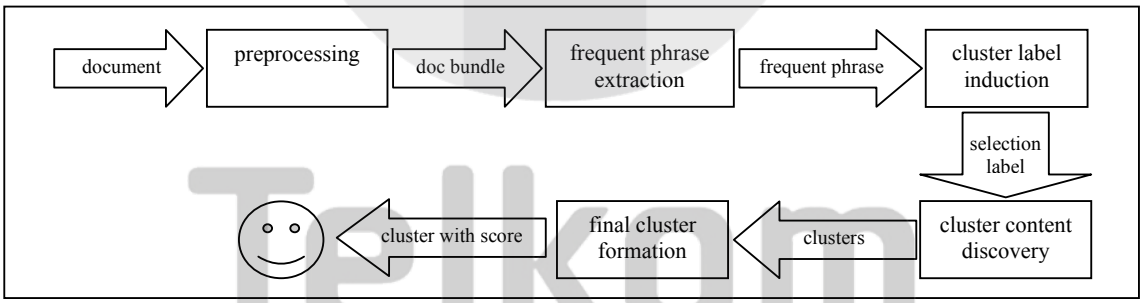
2. Kata-kata bahasa asing dalam isi dokumen akan dianggap sebagai kata dalam bahasa Indonesia.
3. Kesalahan ketik suatu kata dianggap sebagai suatu kata yang baru.
4. *Sample* dokumen yang dipakai adalah dokumen abstraksi Tugas Akhir dan Proyek Akhir jurusan teknik informatika yang terdapat pada perpustakaan ITTelkom.

1.3 Tujuan

Tujuan penulisan yang ingin dicapai dalam tugas akhir ini adalah:
1. Menerapkan algoritma LINGO pada aplikasi pengelompok hasil pencarian.
2. Menganalisis performansi algoritma LINGO terhadap kesesuaian penamaan suatu *cluster* dengan isi dokumen-dokumen di dalam *cluster* tersebut, dengan menghitung *relevancy*-nya dengan menghitung *precision* dan *recall*-nya.

1.4 Metodologi penyelesaian masalah

Metodologi pembahasan yang digunakan dalam penelitian Tugas Akhir ini adalah:
1. Mengumpulkan bahan-bahan referensi yang akan menunjang proses penelitian, seperti jurnal-jurnal, artikel-artikel, paper tentang *search engine* dan algoritma LINGO.
2. Studi Literatur tentang *search engine* dan algoritma LINGO.
3. Penerapan algoritma LINGO pada perangkat lunak pengelompok hasil pencarian, dengan langkah-langkah sebagai berikut:
    a. Preprocessing, stemming dan stopword dikenakan pada dokumen.
    b. Frequent Phrase Extraction, menentukan frase-frase untuk kandidat penamaan klaster.
    c. Cluster label induction, melakukan induksi terhadap hasil pada langkah b.
    d. Cluster content discovery, menerapkan Vector Space Model terhadap dokumen.
    e. Final cluster formation, memberikan peringkat terhadap klaster yang dihasilkan.



Gambar 1-1: Gambaran Proses LINGO

4. Analisis performansi algoritma LINGO terhadap kesesuaian penamaan klaster dengan isi dokumen-dokumen yang merupakan anggotanya.
5. Membuat kesimpulan dari hasil penelitian.

5. Kesimpulan

5.1 Kesimpulan

Berdasarkan pengujian dan analisis yang telah dibahas dan dilaksanakan pada bab tiga dan empat, maka dapat disimpulkan beberapa hal sebagai berikut:

1. Dokumen koleksi abstrak Tugas Akhir dan Proyek Akhir bisa dikelompokkan kedalam kelompok-kelompok tematik (*cluster*) dengan menggunakan algoritma LINGO.
2. Pada pengujian menggunakan metode TF-IDF, nilai *similarity* dokumen yang tidak sama dengan nol terhadap label klaster cenderung memenuhi *snippet assignment threshold*.
3. Pada pengujian dengan metode TF hasil *term* label klaster yang dihasilkan lebih beragam dari pada dengan menggunakan metode TF-IDF.
4. Metode pembobotan TF dan TF-IDF pada perhitungan *precision* dan *recall* menghasilkan nilai yang beragam. Tidak ada salah satu metode dengan nilai selalu lebih bagus dibandingkan dengan metode yang lain.

5.2 Saran

Sebagai acuan dalam melengkapi atau memperbaiki hasil analisis data yang dilakukan dalam tugas akhir ini. Ada beberapa saran yang dapat dijadikan pertimbangan bagi analisis data selanjutnya, diantaranya :

1. Aplikasi yang dibuat tidak hanya terbatas untuk dokumen berbahasa Indonesia.
2. Aplikasi hendaknya dilengkapi dengan deteksi bahasa dari dokumen yang diproses.
3. Label cluster yang dihasilkan program ini bisa lebih baik jika memiliki proses *preprocessing* yang lebih baik pula.

Daftar Pustaka

[1] Adriani, Mirna. *Stemming Indonesian: A Confix-Stripping Approach*. 2006. University of Indonesia. Indonesia.

[2] Arifin, Agus Zainal. *Penggunaan Digital Tree Hibrida pada Aplikasi Information Retrieval untuk Dokumen Berita*. Jurusan Teknik Informatika, FTIF, Institut Teknologi Sepuluh Nopember. Surabaya. Indonesia.

[3] Asian, Jelita. *A Testbed for Indonesian Text Retrieval.* School of Computer Science and Information Technology RMIT University, Melbourne, Australia.

[4] Asian, Jelita. *Stemming Indonesian*. School of Computer Science and Information Technology RMIT University, Melbourne, Australia.

[5] Ayuningtias, Vidya. *Pengkategorian Hasil Pencarian Dokumen dengan Clustering*. 2007. ITTelkom, Bandung. Indonesia.

[6] Baeza, Ricardo. *Modern Information Retrieval*. Januari 1999. ACM Press, New York.

[7] Budhi, Gregorius S. ALGORITMA *PORTER STEMMER FOR BAHASA INDONESIA* UNTUK *PRE-PROCESSING TEXT MINING* BERBASIS METODE *MARKET BASKET ANALYSIS.* Petra Jurusan Teknik Informatika. Indonesia.

[8] Firdaus, Yanuar. *Diktat Kuliah Information Retrieval*. 2007. ITTelkom, Bandung.

[9] Geraci, Filippo. *A Scalable Algorithm for HighQuality Clustering of Web Snippets.* Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche Via G Moruzzi, Pisa, Italy.

[10] Geraci, Filippo. *Cluster Generation and Cluster Labelling for Web Snippets: a Fast and Accurate Hierarchical Solution.* Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche, Via G Moruzzi 1, Pisa, Italy.

[11] Mecca Giansalvatore. *A New Algorithm for Clustering Search Results.* Dipartimento di Matematica e Informatica Università della Basilicata Potenza. Italy.

[12] Osinski, Stanislaw. *Conceptual Clustering Using Lingo Algorithm: Evaluation on Open Directory Project Data*. 2004. Institute of Computing Science, Pozna´n University of Technology, Poland.

[13] Osinski, Stanislaw. *DIMENSIONALITY REDUCTION TECHNIQUES FOR SEARCH RESULTS CLUSTERING.* Department of Computer Science The University of Sheffield, UK.

[14] Osinski, Stanislaw. *Improving Quality of Search Results Clustering with Approximate Matrix Factorisations.* Poznan Supercomputing and Networking Center, Poznan, Poland.

[15] Osinski, Stanislaw. *LINGO: Search Results Clustering Algorithm Based on Singular Value Decomposition*. 2003. Institute of Computing Science, Pozna´n University of Technology, Poland.

[16] Passadore, Andrea. *AN INDEXING AND CLUSTERING ARCHITECTURE TO SUPPORT DOCUMENT RETRIEVAL IN THE MAINTENANCE SECTOR.* Via San Nazaro Genova, Italy

[17] Tala, Fadillah Z. *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*. Institute for Logic, Language and Computation Universiteit van Amsterdam. Netherlands

[18] Total.or.id. Maret 2009.

[19] Wikipedia.com/, 2009.

[20] Weiss, Dawid. *Descriptive Clustering as a Method for Exploring Text Collection.* Pozna´n University of Technology Institute of Computing Science, Poland.

[21] WRÓBLEWSKI, MICHAŁ. *A HIERARCHICAL WWW PAGES CLUSTERING ALGORITHM BASED ON THE VECTOR SPACE MODEL*. Poznań University of Technology, Poland