IMPLEMENTASI DAN ANALISIS WEB SEARCH RESULTS CLUSTERING DENGAN ALGORITMA SEMANTIC HIERARCHICAL ONLINE CLUSTERING (SHOC)

Isa Albanna Susianto[1], Yanuar Firdaus A.w.[2], Kusuma Ayu Laksitowening[3]

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

---

Abstrak

Clustering (pengelompokan dokumen) merupakan salah satu teknik yang dapat digunakan untuk memudahkan user dalam menemukan dokumen web yang diinginkan dari sejumlah retrieved documents yang dihasilkan search engine. Teknik ini mengelompokkan dokumen berdasarkan kategori tertentu, sehingga penelusuran user terhadap retrieved documents akan lebih mengerucut.

Algoritma Semantic Hierarchical Online Clustering (SHOC) merupakan salah satu algoritma clustering yang mengelompokkan dokumen web hasil pencarian ke dalam cluster tertentu berdasarkan frase-frase kunci yang terdapat dalam dokumen tersebut. Tugas Akhir ini mengimplementasi dan menganalisis clustering hasil pencarian search engine dengan menggunakan algoritma SHOC.

Hasil pengujian menunjukkan bahwa algoritma SHOC mampu memisahkan retrieved documents yang relevan dan tidak dengan performansi yang dipengaruhi oleh kualitas hasil pencarian dan karakteristik dokumen. Algoritma SHOC akan optimal untuk mengelompokkan dokumen-dokumen yang saling berbagi frase kunci. Dan untuk pengaruh kualitas search engine, precision yang terlalu kecil akan menyebabkan banyaknya cluster "sampah" terbentuk, sedangkan recall yang terlalu kecil akan mengurangi ketepatan pembentukan cluster. Untuk menangani kualitas search engine yang kurang baik, nilai threshold cluster quality pada algoritma SHOC perlu diset sesuai dengan kualitas search engine, sehingga dokumen yang relevan bisa tetap dikelompokkan.

Kata Kunci : search engine, retrieved documents, clustering, frase kunci, algoritma SHOC

Abstract

Document clustering is a technique that can be implemented to ease user in finding the expected documents from many documents retrieved by the search engine. This technique groups documents based on certain category, so that the user's searching on retrieved documents can be simplified.

Semantic Hierarchical Online Clustering (SHOC) algorithm is one of clustering algorithms which groups web documents into certain cluster based on key phrases contained in the documents. This research implements and analyzes web search result clustering using SHOC algorithm.

The testing result shows that SHOC algorithm is able to separate the relevant and unrelevant retrieved documents where the performance depends on the quality of the search engine and the documents' characteristic. SHOC algorithm is suitable for grouping documents that share key phrases each other. And for the influence of the quality of search engine, poor search engine's precision will generate many "waste clusters", and poor search engine's recall will decrease the accuracy of the generated clusters. To handle the poorness of the search engine, the value of cluster quality threshold in SHOC algorithm needs to be set according to the quality of search engine, so the relevant documents can still be grouped.

Keywords : search engine, retrieved documents, clustering, key phrases, SHOC algorithm

---

1. Pendahuluan

1.1 Latar belakang

Ketika *user* melakukan pencarian dengan menggunakan suatu *term* pada *search engine*, sering kali *user* mendapatkan sangat banyak dokumen hasil pencarian (*retrieved documents*). Dari *retrieved documents* tersebut, tidak semua mengacu pada hasil yang diinginkan. Bahkan mungkin saja hasil yang diinginkan bukan terdapat pada dokumen hasil pencarian yang pertama, kedua, atau ketiga, melainkan pada dokumen di bagian akhir pencarian, sehingga mau tidak mau *user* dipaksa menelusuri satu per satu *retrieved documents* dari awal sampai akhir untuk mendapatkan dokumen yang diinginkan.

Salah satu solusi untuk masalah ini adalah dengan mengelompokkan *retrieved documents* ke dalam kelompok-kelompok tertentu. Pengelompokan dilakukan berdasarkan kemiripan topik yang dibahas dalam suatu dokumen, sehingga *user* hanya perlu mengakses kelompok dokumen yang memuat topik sesuai dengan topik yang diinginkan, bukannya memeriksa isi *retrieved documents* satu per satu. Teknik untuk mengelompokkan dokumen ini disebut *document clustering*. *Document clustering* juga bisa diterapkan dalam mengelompokkan hasil pencarian *search engine*, sehingga dapat menyelesaikan permasalahan di atas.

Banyak metode atau algoritma dapat diterapkan dalam *document clustering*, salah satunya adalah algoritma *Semantic Hierarchical Online Clustering* (SHOC). Algoritma ini merupakan algoritma yang berbasis pada frase kunci yang ada dalam dokumen. Awalnya, algoritma SHOC diterapkan untuk bahasa dengan jumlah alfabet yang banyak seperti bahasa Mandarin, namun dalam Tugas Akhir ini diterapkan algoritma SHOC pada bahasa Inggris, dengan kemudian menganalisis performansinya berdasarkan dua parameter kualitas *clustering* yakni F-Measure dan Dunn Index [4]. F-Measure menunjukkan ketepatan pembagian dokumen ke dalam *cluster*, sedangkan Dunn Index mengukur jarak dokumen intra dan inter *cluster*. Sebelumnya, belum pernah diterapkan analisis dengan kedua parameter tersebut pada algoritma SHOC.

1.2 Perumusan masalah

Mengacu pada latar belakang di atas, ada beberapa masalah yang akan diselesaikan dalam Tugas Akhir ini, diantaranya yaitu :

1. Bagaimana menerapkan *web search results clustering* dengan algoritma SHOC?
2. Sejauh mana performansi algortitma SHOC dalam membagi dokumen hasil pencarian ke dalam *cluster* yang tepat?
3. Bagaimana analisis terhadap kondisi yang mempengaruhi kualitas *web search results clustering* dengan algoritma SHOC.

1.3 Batasan masalah

Berikut adalah batasan masalah pada Tugas Akhir ini :

1. *Document collection* yang dipakai merupakan dokumen teks bahasa Inggris dalam bentuk file HTML (selain HTML 5).
2. *Document collection* merupakan dokumen teks bertopik yang biasa digunakan untuk pengujian *search engine* yang didapat dari ftp://ftp.cs.cornell.edu/pub/smart/med.
3. Pengujian dilakukan secara *offline*.

1.4 Tujuan

Berikut adalah tujuan pengerjaan Tugas Akhir ini :

1. Menerapkan algoritma SHOC untuk mengelompokkan hasil pencarian *search engine*.
2. Menganalisis ketepatan algoritma SHOC dalam mengelompokkan dokumen ke dalam *cluster* tertentu. Parameter analisis yang digunakan adalah F-Measure (berdasarkan *precision* dan *recall clustering*), Dunn Index (berdasarkan *distance measure* dokumen yang dikelompokkan).
3. Menganalisis kondisi yang mempengaruhi performansi *web search results clustering* dengan algoritma SHOC dan bagaimana kondisi optimalnya.

1.5 Metodologi penyelesaian masalah

Berikut adalah metodologi yang digunakan dalam penyelesaian masalah dalam Tugas Akhir ini :

1. Studi literatur.
   Pencarian materi-materi dan referensi yang berkaitan dengan permasalahan yang dibahas, seperti materi tentang *information retrieval*, *document clustering*, algoritma SHOC, dan lain-lain.
2. Analisis dan perancangan kebutuhan sistem.
   Merupakan tahap perancangan sistem yang akan dibuat, yakni sebuah *search engine* yang menerapkan *document clustering* dengan algoritma SHOC, serta membuat analisis terhadap kebutuhan sistem yang akan dibuat.
3. Implementasi sistem.
   Mengimplementasikan sistem ke dalam program untuk kemudian diuji dan dianalisis. Proses implementasi dilakukan dengan tahapan :
   a. mengumpulkan *document collection*,
   b. membangun sebuah *search engine*, dan
   c. menerapkan algoritma SHOC pada *post-processing search engine* tersebut untuk mengelompokkan hasil pencarian.
4. Pengujian sistem.
   Menguji sistem yang telah diimplementasikan dan menganalisis hasil performansi dari *search engine* yang menerapkan *document clustering* dengan algoritma SHOC. Performansi dilihat dari waktu *clustering* dan ketepatan pengelompokan dokumen ke dalam *cluster*s. Pengujian dilakukan dengan memasukkan kata kunci tertentu pada *search engine*,

kemudian mengukur waktu *clustering* (waktu pengelompokkan dokumen, tidak termasuk waktu pencarian), dan melihat hasil akhir pengelompokkan *cluster* untuk kemudian dicocokkan dengan pengelompokkan yang diharapkan (*clustering judgement*). Pengujian dilakukan berkali-kali dengan kata kunci yang berbeda-beda.

5. Analisis hasil pengujian dan pengambilan keputusan.
   Analisis terhadap hasil pengujian sistem dan perumusan kesimpulan terhadap hasil analisis. Aspek yang dianalisis adalah waktu *clustering* untuk melihat efisiensi algoritma SHOC, serta ketepatan *clustering* untuk melihat efektifitas dari algoritma SHOC. Ketepatan *clustering* dilihat dari parameter F-Measure dan Dunn Index.
6. Penyusunan laporan Tugas Akhir.
   Penyusunan laporan semua tahap yang telah dilakukan mulai dari tahap studi literatur sampai perumusan kesimpulan.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan hasil pengujian dan analisis yang telah dilakukan, maka dapat diambil kesimpulan berikut ini:

1. Algoritma SHOC terbukti mampu mengelompokkan dokumen relevan ke dalam satu *cluster*, ditandai dengan sebandingnya nilai F-Measure clustering terhadap *recall search engine*. Selain itu, dengan menggunakan *search engine* yang rata-rata nilai *precision*-nya sekitar 0.50 dan *recall*-nya sekitar 0.47, algoritma SHOC mampu memisahkan dokumen relevan dan tidak relevan dengan rata-rata akurasi mencapai 77% (dinilai dari selisih persentasi dokumen relevan dan tidak relevan dalam satu *cluster*).
2. Dalam penerapannya untuk mengelompokkan hasil pencarian pada *search engine*, performansi algoritma SHOC dipengaruhi oleh *precision* dan *recall search engine*. *Precision* dan *recall* yang bagus akan menghasilkan kualitas *cluster* yang bagus pula. *Precision* yang terlalu kecil akan menyebabkan banyaknya *cluster* ”sampah” yang terbentuk (nilai Dunn Index-nya menurun), sedangkan *recall* yang terlalu kecil akan menyebabkan menurunnya ketepatan pembentukan *cluster* (nilai F-Measure-nya menurun).
3. Algoritma SHOC mengelompokkan dokumen berdasarkan persebaran frase kunci pada dokumen, sehingga algoritma ini akan mengelompokkan *retrieved documents* yang relevan dengan lebih tepat apabila dokumen-dokumen tersebut saling berbagi frase.
4. Besarnya nilai *threshold cluster quality* perlu diset bergantung pada kualitas *search engine*. Dengan *precision* atau *recall* yang rendah, nilai *threshold* ini perlu diset tinggi, bahkan diset bernilai 1, agar *cluster* yang berisi dokumen yang relevan bisa lebih banyak terambil. Namun semakin tingginya nilai *threshold* ini akan menyebabkan semakin banyak pula *cluster* ”sampah” yang terbentuk. Apabila nilai *precision* dan *recall search engine*-nya tinggi, maka nilai *threshold* ini bisa diset lebih rendah sehingga *cluster* yang berisi dokumen relevan bisa terambil dan juga jumlah *cluster* ”sampah” bisa dikurangi.

5.2 Saran

Setelah tugas akhir ini selesai dilakukan dan dilakukan analisis, penulis memiliki beberapa saran sebagai berikut:

1. Dalam penerapan algoritma SHOC, besarnya nilai *threshold cluster quality* perlu disesuaikan dengan nilai *precision* dan *recall search engine*.
2. Setelah dilakukan *clustering*, sebaiknya dilakukan pengurutan *cluster* berdasarkan relevansi dokumen anggotanya terhadap *query* sehingga *cluster* yang berisi dokumen relevan bisa ditempatkan di urutan pertama dan *cluster* yang berisi dokumen tidak relevan ditempatkan di urutan akhir.

Referensi

[1] Branson, Steve, Ari Greenberg. *Clustering Web Search Results Using Suffix Tree Methods*. Stanford, USA : Stanford University. Diunduh pada: http://www.stanford.edu/class/cs276a/projects/reports/arigreen-sbranson.pdf, 17 September 2009.

[2] Cao, Guihong, Dawei Song, Peter Bruza. 2003. *Suffix Tree Clustering on Post-retrieval Documents*. Brisbane : The University of Queensland. Diunduh pada: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.6.429&rep=rep1&type=pdf, 17 September 2009.

[3] Dell, Zhang, Dong Yisheng. *Semantic, Hierarchical, Online Clustering of Web Search Results*. Georgia : 3rd International Workshop on Web Information and Data Management. Diunduh pada: http://www.dcs.bbk.ac.uk/~dell/publications/dellzhang_shoc.pdf, 4 Oktober 2009.

[4] Eissen, Sven Meyer zu, Benno Stein. 2002. *Analysis of Clustering Algorithms for Web-based Search*. Paderborn : Paderborn University. Diunduh pada: http://www.uni-weimar.de/medien/webis/publications/downloads/papers/stein_2002e.pdf, 29 September 2009.

[5] Fung, Benjamin C. M., dkk. *Hierarchical Document Clustering*. Burnaby : Simon Fraser University. Diunduh pada: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.119.2558&rep=rep1&type=pdf, 4 Oktober 2009.

[6] Kummamuru, Krishna, dkk. *A Hierarchical Monothetic Document Clustering Algorithm for Summarization and Browsing Search Results*. New Delhi : IBM India Research Lab. Diunduh pada: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1.8184&rep=rep1&type=pdf, 17 September 2009.

[7] Oren, Zamir, Etzioni Oren. 1998. *Web Document Clustering: A Feasibility Demonstration*. Seattle : University of Washington. Diunduh pada: http://www.cs.washington.edu/research/projects/WebWare1/etzioni/www/papers/sigir98.pdf, 17 September 2009.

[8] Oren, Zamir, Etzioni Oren. 1999. *Grouper: A Dynamic Clustering Interface to Web Search Results*. Seattle : University of Washington. Diunduh pada: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.31.8216&rep=rep1&type=pdf, 12 September 2009.

[9] Oren, Zamir, Etzioni Oren. *Web Document Clustering*. Seattle : University of Washington. Diunduh pada: http://www.sigmod.org/disc/disc99/disc/dmkd/ze28.pdf, 18 September 2009.

[10] Osinski, Stanislaw. 2003. *An Algorithm for Clustering of Web Search Result*. Poznan : Poznan University of Technology. Diunduh pada: http://project.carrot2.org/publications/osinski-2003-lingo.pdf, 29 September 2009

[11] Osinski, Stanislaw. 2004. *Dimensionality Reduction Techniques for Search Results Clustering*. Sheffield : University of Sheffield. Diunduh pada: http://project.carrot2.org/publications/osinski04-dimensionality.pdf, 4 Oktober 2009.

[12] Steinbach, Michael, George Karypis, Kumar Vipin. *A Comparison of Document Clustering Techniques*. Minnesota : University of Minnesota. Diunduh pada: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.134.1422&rep=rep1&type=pdf, pada 20 September 2009.