SISTEM DETEKSI KEMIRIPAN DOKUMEN DENGAN ALGORITMA *COSINE SIMILARITY* DAN *SINGLE PASS CLUSTERING*

Sugiyamta

Abstrak

Kesamaan Dokumen dapat digunakan untuk menjadi petunjuk dan contoh mencari informasi yang sama. Kemampuan mencari kesamaan ini dapat mengurangi waktu. Untuk menggambarkan tingkat kesamaan antara dokumen dapat diukur oleh Metode Cosine Similarity. Berdasarkan tingkat kesamaan dokumen dapat diklasifikasikan dengan menggunakan Algoritma Single Pass Clustering. Untuk mendeteksi tingkat kemiripan dari objek tersebut dibuat Aplikasi dengan menggunakan bahasa pemrograman PHP, web server Apache 2.0, dan MySql 5 database server. Objek penelitian ini adalah tentang Deteksi Kesamaan Abstraksi dari Skripsi Mahasiswa. Hasil Aplikasi ini untuk menyajikan rasio kesamaan sistematis dan juga manual yang menghasilkan tingkat kesamaan, sehingga sistem ini daoat memberikan informasi tentang kesamaan dokumen abstrak skripsi mahasiswa karena akurasinya mencapai 99%.

Keywords: A*bstrak skripsi, Dokumen Similarity, pencarian informasi, Cosine Similarity, Single Pass Clustering.*

1. Pendahuluan

Kemiripan dokumen (*document similarity*) dapat digunakan sebagai alat pencarian informasi lain yang sejenis, sehingga dapat mempersingkat waktu. Kemampuan pencarian kemiripan dokumen biasanya diimplementasikan pada sebuah artikel berita dan jurnal (Strasberg, 2002).

Metode *cosine similarity* merupakan metode yang digunakan untuk menghitung tingkat kesamaan (similarity) antar dua buah dokumen.

Klastering (*clustering*) didefinisikan sebagai upaya mengelompokkan data ke dalam klaster sedemikian sehingga data-data di dalam klaster yang sama lebih memiliki kesamaan dibandingkan dengan data-data pada klaster yang berbeda.

Metode Single Pass Clustering adalah metode yang menggunakan strategi disain Bottom-Up yang dimulai dengan meletakkan setiap obyek sebagai sebuah klaster tersendiri dan selanjutnya menggabungkan klaster tersebut menjadi klaster yang lebih besar dan lebih besar lagi sampai akhirnya semua obyek menyatu dalam sebuah klaster atau proses dapat berhenti jika telah mencapai batasan kondisi tertentu.

Tujuan dari penelitian ini adalah membuat sistem deteksi kemiripan dokumen (document similarity) menggunakan Algoritma Cosine Similarity dan teknik mengelompokkan dokumen dengan Algoritma Single Pass Clustering..

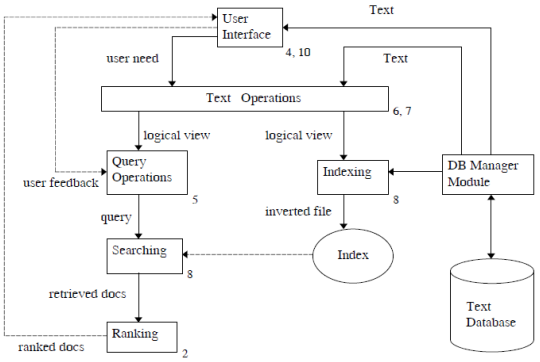
2. Telaah Pustaka

2.1. Sistem Temu Kembali Informasi

Sistem Temu Kembali Informasi merupakan suatu system yang menyimpan informasi dan menemukan kembali informasi tersebut. Secara konsep bahwa ada beberapa dokumen atau kumpulan record yang berisi informasi yang diorganisasikan ke dalam sebuah media penyimpanan untuk tujuan mempermudah ditemukan kembali. Dokumen yang tersimpan tersebut dapat berupa kumpulan record informasi bibliografi maupun data lainnya (Salton 1989).

Teknik pencarian informasi pada Sistem Temu Kembali Informasi (Information Retrieval System) berbeda dengan sistem pencarian pada sistem manajemen basis data (DBMS). Dalam sistem temu kembali terdapat dua bagian utama yaitu bagian pengindeksan (indexing) dan pencarian (searching). Kedua bagian tersebut memiliki peran penting dalam proses temu kembali

informasi. Gambar 1 menggambarkan proses temu kembali informasi.



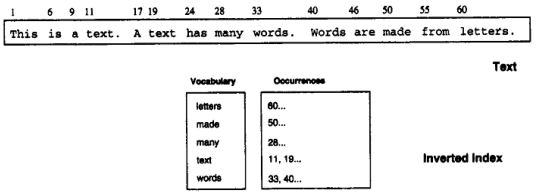
Gambar 1 Proses Temu Kembali Informasi (Baeza dan Ribeiro, 1999)

2.2. Klastering Dokumen

Klastering biasa digunakan pada banyak bidang, seperti : data mining, pengenalan pola (pattern recognition), pengklasifikasian gambar (image classification), ilmu biologi, pemasaran, perencanaan kota, pencarian dokumen, dan lain sebagainya. Tujuan dari klastering adalah untuk menentukan pengelompokan dari suatu set data. Akan tetapi tidak ada ”ukuran terbaik” untuk pengelompokan data. Untuk pengelompokkan data tergantung tujuan akhir dari klastering, maka diperlukan suatu kriteria sehingga hasil klastering seperti yang diinginkan.

2.3. Index Inverted

Inverted file atau index inverted adalah mekanisme untuk pengindeksan kata dari koleksi teks yang digunakan untuk mempercepat proses pencarian. Struktur inverted file terdiri dari dua elemen, yaitu: kata (vocabulary) dan kemunculan (occurences). Kata-kata tersebut adalah himpunan dari kata-kata yang ada pada teks, atau merupakan ekstraksi dari kumpulan teks yang ada, dan tiap kata terdapat juga informasi mengenai semua posisi kemunculannya (occurences) secara rinci. Posisi dapat merefer kepada posisi kata ataupun karakter. Hal ini dapat dilihat dengan jelas dengan memperhatikan Gambar 2

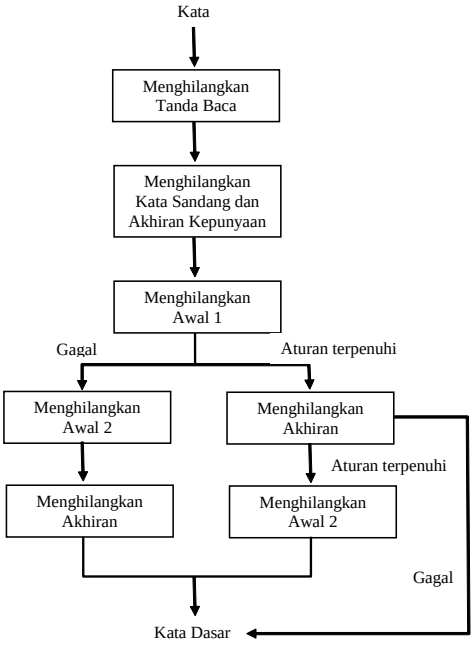


Gambar 2 Contoh Teks dan Inverted File-nya

2.4. Stemming

Stemming merupakan suatu proses untuk menemukan kata dasar dari sebuah kata. Dengan menghilangkan semua imbuhan (affixes) baik yang terdiri dari awalan (prefixes), sisipan (infixes), akhiran (suffixes) dan confixes (kombinasi dari awalan dan akhiran) pada kata turunan. Stemming digunakan untuk mengganti bentuk dari suatu kata menjadi kata dasar dari kata tersebut yang sesuai dengan struktur morfologi Bahasa Indonesia yang baik dan benar.

Arsitektur proses stemming untuk bahasa Indonesia dapat dilihat pada Gambar 3



Gambar 3 Proses Stemming Algoritma Tala (Tala, 2003)

2.5. Cosine Similarity

Cosine similarity merupakan metode yang digunakan untuk menghitug tingkat kesamaan (similarity) antar dua buah objek. Untuk tujuan klastering dokumen, fungsi yang baik adalah fungsi cosine similarity
Untuk notasi himpunan digunakan rumus :

$$Similarity(X,Y)=\frac{|X\cap Y|}{|X|^{\frac{1}{2}}.|Y|^{\frac{1}{2}}} \quad \text{.....(1)}$$

Dimana :

$|X\cap Y|$ adalah jumlah term yang ada pada dokumen *X* dan yang ada pada dokumen *Y*

$|X|$ adalah jumlah term yang ada pada dokumen *X*

$|Y|$ adalah jumlah term yang ada pada dokumen *Y*

2.6. Metode Single Pass Clustering

Single Pass Clustering merupakan suatu tipe klastering yang berusaha melakukan pengelompokan data satu demi satu dan pembentukan kelompok dilakukan seiring dengan pengevaluasian setiap data yang dimasukkan ke dalam proses klaster. Pengevaluasian tingkat kesamaan antar data dan klaster dilakukan dengan berbagai macam cara termasuk menggunakan fungsi jarak, vectors similarity, dan lain-lain.
Algoritma yang sering digunakan dalam Single Pass Clustering adalah, unt masing-masing data d :

1) loop
   a) menemukan a klaster c yang memaksimalkan an fungsi objektif
   b) jika nilai dari fungsi subjektif > a maka nilai ambang masukkan d didalam c
   c) jika a klaster baru maka a adalah hanya data d
2) akhir loop

Dalam menggunakan algoritma ini, dua hal yang perlu menjadi perhatian adalah penentuan objective function dan penentuan threshold value. Objective function yang ditentukan haruslah sebisa mungkin mencerminkan keadaan data yang dimodel dan dapat memberikan nilai tingkat kesamaan atau perbedaan yang terkandung di dalam data tersebut. Penentuan threshold value juga merupakan hal yang subjektif, makin besar nilai threshold, makin mudah suatu data untuk bergabung ke dalam suatu klaster, dan demikian juga sebaliknya. (Klampanos, 2006).

3. Metodologi Penelitian

3.1. Bahan Penelitian

Bahan yang digunakan dalam penelitian ini berupa data Judul dan Abstrak Skripsi sebanyak 550 dokumen.

3.2. Alat Penelitian

Alat penelitian yang digunakan dalam sistem deteksi kemiripan dokumen adalah sebagai berikut :

1. Spesifikasi perangkat keras yang digunakan processor Intel Core i5, 2.30 GHz, RAM 4 GB, Hardisk 360 GB, monitor.
2. Perangkat lunak yang digunakan dalam sistem ini adalah Microsoft Windows 7 Ultimate, Bahasa pemrograman PHP, webserver Apache 2.0 dan server basis data MySql 5.

3.3. Deskripsi Sistem

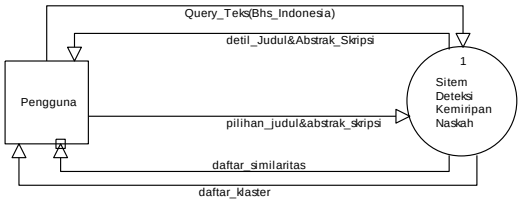
Sistem ini adalah sistem yang digunakan untuk memberikan kemampuan pencarian kemiripan dokumen pada basis data Judul dan Abstrak Skripsi. Basis data yang dimaksud adalah data Judul dan Abstrak Skripsi dengan tujuan agar mempunyai kemampuan untuk mencari dokumen-dokumen yang mirip antara satu dengan lainnya. Sehingga pengguna dapat mencari, membaca dan mengelola dokumen Judul dan Abstrak Skripsi dengan lebih mudah.

3.4. Diagram Alir Dokumen

Diagram alir data merupakan gambaran secara grafis dari alir data yang melewati sistem. Diagram alir data pada penelitian ini dirancang diagram konteks dan diagram level 1.

Diagram kontek sistem deteksi kemiripan dokumen diperlihatkan dalam

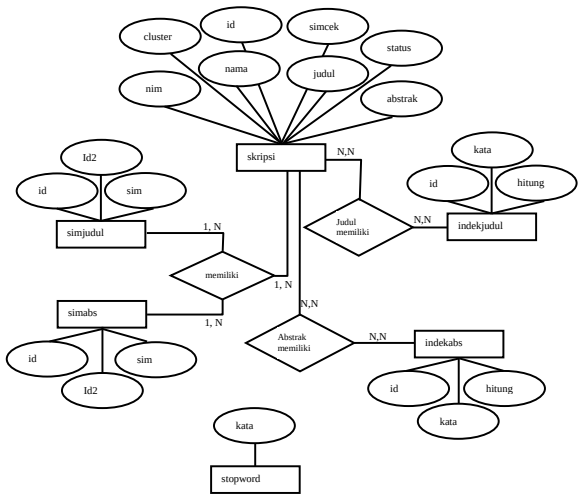
Gambar 4. Sistem ini memiliki satu entitas luar yaitu pengguna. Entitas luar pengguna memberikan msukan query (teks bahasa Indonesia) dan sistem akan mengeluarkan keluaran daftar judul dan abstrak berdasarkan rangking bobot kata masukan query terhadap judul dan abstrak. Pengguna dapat memilih salah satu judul dan abstrak untuk dibaca, maka sistem akan menampilkan judul dan abstrak dan daftar judul dan abstrak yang mirip dengan judul dan abstrak tersebut.



Gambar 4 Diagram konteks

3.5. Perancangan Basis Data

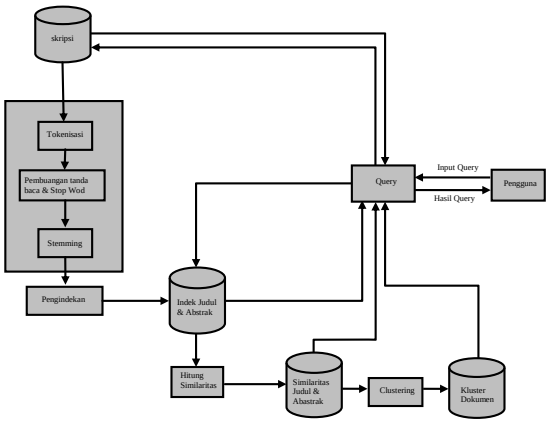
Rancangan sistem basis data pada sistem deteksi kemiripan dokumen dan clustering menggunakan Entity Relationshyp Diagram yang diperlihatkan pada Gambar 5. Gambar tersebut memperlihatkan tabel skripsi yang mempunyai hubungan one to many dengan tabel simjudul dan simabs. Tabel skripsi juga mempunyai hubungan one to many dengan tabel indekjudul dan indekabs. Tabel IndekJudul dan Indekabs mempunyai hubungan many to many dengan tabel skripsi. Sedangkan tabel stopword tidak mempunyai hubungan dengan tabel yang lainnya.



Gambar 5 E-R diagram

3.6. Arsitektur Sistem

Deteksi kesamaan dokumen dengan algoritma Single Pass Clustering sebagai suatu sistem memiliki beberapa proses (modul) yang membangun sistem secara keseluruhan. Modul Sistem Deteksi kesamaan dokumen terdiri dari : modul tokenisasi (tokenizations), modul pembuangan stop word (stop word removal), modul pengubahan kata dasar (stemming), modul pengindeksan kata (term indexing), kesamaan kata (term similarity) dan modul pengelompokan (clustering). Secara lengkap rancangan dari modul Deteksi kesamaan dokumen dapat dilihat pada Gambar 6



Gambar 6 Rancangan Sistem Deteksi Kesamaan Dokumen

4. Hasil dan Pembahasan

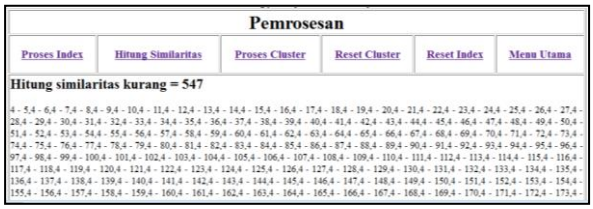
4.1. Hasil Penelitian

Proses stemming adalah proses mengembalikan kata dalam bentuk dasarnya, proses ini merupakan bagian dari proses filtering. Pada penelitian ini menggunakan algoritma Tala (2003).

Proses Pengindekan, setelah sebuah kata melalui proses stemming maka kata tersebut akan disimpan dalam master table IndekJudul dan IndekAbs (indek abstrak), yang berfungsi mencatat kemunculan suatu kata pada seluruh koleksi abstrak skripsi. Kata terlebih dahulu dicocokan keberadaannya pada table per id skripsi, apabila terdapat kata yang sama pada table indek maka kata tersebut hanya di hitung 1, kemudian kata yang lain disisipkan

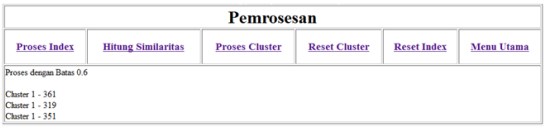
record baru pada Tabel IndekJudul atau Tabel Indekabs (indek abstrak). Jadi kata yang sama dalam id yang sama hanya tersimpan satu kali.

Penghitungan Nilai Cosine Similarity adalah proses penghitungan kesamaan (similarity). Proses ini adalah menghitung kesamaan atar dokumen yang dihitung berdasarkan id yang terdapat pada tebel skripsi dengan membandingkan id 1 dengan id 2, id 1 dengan id 3, id 1 sampai dengan id 550, sampai dengan id 550 terakhir yaitu id 549 dengan id 550, karena didalam penelitian ini digunakan 550 dokumen. Proses similaritas Judul dan Similaritas abstrak terlihat pada gambar 7



Gambar 7 Proses Hutung Similaritas Judul dan Abstrak

Klaster Dokumen dari proses hitung similaritas akan digunakan untuk melakukan proses klaster. Proses klastering dilakukan dengan mencari nilai kemiripan yang maksimum dari tabel similaritas. dokumen pada id 301 mempunyai kemiripan dengan id 351 dengan nilai tertinggi yaitu 1 atau 100% maka dikumen dengan id 301 dan id 351 menjadi kandidat dari klaster pertama, dokumen pada id 301 mempunyai kemiripan dengan id 351 dengan nilai 1 atau 100%. Kandidat klaster kedua pada id 299 dengan id 372 yang mempunyai nilai kemiripan 0.9730124597112 atau 97,30%. Proses akan dilanjutkan dengan mencari kandidat untuk klaster yang lain. Implementasi proses klaster akan dilakukan tahap yang sama sampai dokumen dalam tabel skripsi selesai dibuat klaster. Proses klaster terlihat pada Gambar. 8



Gambar 8 Proses Klastering dengan Threshold 0.6

4.2. Pembahasan

Pengukuran proses stemming, dari jumlah dokumen 550 abstrak skripsi, setelah dilakukan proses stemming, sehingga terbentuk sebanyak 4106 kata yang digunakan pada abstrak skripsi, kata-kata tersebut telah bebas dari stopword.

Setelah dilakukan evaluasi, ada bebarapa kesalahan yaitu karena overstemming, hal ini paling banyak terjadi, untuk bahasa asing tidak terjadi perubahan karena akhiran dan awalan tidak dikenali oleh sistem dan kesalahan terjadi pada nama orang/istilah/singkatan.Pada Tabel 1 merupakan contoh evaluasi kesalahan proses stemming.

Tabel 1 Hasil Evaluasi Kesalahan Stemming

Jenis Kesalahan	Contoh	Hasil stemmer	Seharusnya
Nama Orang, tempat, istilah, singkatan	Kariadi, determinasi, kaprogdi	Kariad, determinas, kaprogd	-tetap-
Bahasa Asing	Application, collection, inventory	-tetap-	-sudah benar-
Kata terlalu banyak dipotong (overstemming)	Metode, bayi, pemberian	Tode, bay, ian	Metode, bayi, beri
Kata terlalu sedikit dipotong (understemming)	Dipengaruhi, keterlambatan	Ngaruh, terlambat	Aruh, lambat

Pengukuran proses similaritas, perhitungan nilai similaritas yang dilakukan oleh sistem adalah sebanyak 150.975 kali proses, hal ini sesuai dengan jumlah record yang terdapat pada Tabel Simabs (similaritas abstrak). Distribusi nilai similaritas dapat diperlihatkan pada Tabel 2.

Tabel 2 Distribusi nilai similaritas

No.	Similaritas	Jumlah record
1	91% - 100%	12
2	81% - 90%	19
3	71% - 80%	22
4	61% - 70%	41
5	51% - 60%	60
6	41% - 50%	155
7	31% - 40%	1.526
8	21% - 30%	19.057
9	11% - 20%	80.227
10	1% - 10%	48.725
11	0%	1.131
Total :		**150.975**

Hasil Similaritas, dibuktikan secara manual dengan membandingkan group kata yang dihasilkan oleh sistem dan group kata yang terdapat pada dokumen abstrak asli, dan kata-kata tersebut telah bebas dari stopword. Sebagai contoh validasi hasil similaritas yaitu id 204 dengan id 240 yang mempunyai nilai kemiripan 0.9256514468702. Tabel 3 merupakan ilustrasi perbandingan validasi hasil similaritas dengan cara manual.

Tabel 3 Perbandingan validasi hasil similaritas

Id	Jumlah kata	Kata sama	Kata tidak sama	Perhitungan Manual
204	61	56	5	56/61 = 0.9180327868852
240	60	56	4	56/60 = 0.9333333333333
hasil perhitungan manual				0.9256830601093

Hasil rata-rata perhitungan dengan cara manual adalah : = (0.9180327868852 + 0.9333333333333) / 2 = **0.9256830601093**

Hasil perhitungan sistem adalah : **0.9256514468702**

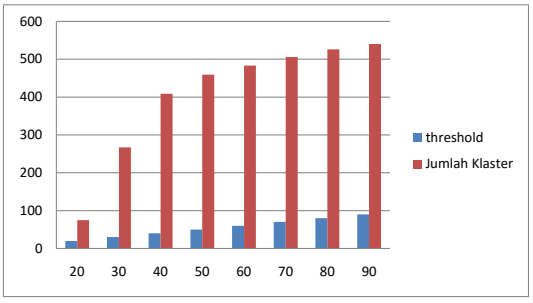
Dari kedua hasil tersebut diatas dapat disimpulkan bahwa, perbandingan nilai similaritas secara sistem dan secara manual menghasilkan nilai yang sama. Sehingga sistem menghasilkan keputusan yang sangat baik, dengan tingkat akurasi 99%.

Hasil Klaster, evaluasi secara manual kesamaan diantara dokumen dalam klaster-klaster yang telah dikelompokkan, yang dihasilkan dari 8 percobaan yang dilakukan dapat dilihat pada tabel 4. Percobaan pengujian dengan nilai batas threshold yang digunakan mulai dari nilai batas threshold 0.2, sampai dengan mendapatkan banyaknya dokumen pada klaster sama dengan jumlah dokumen.

Tabel 4 Hasil Jumlah Klaster

No	threshold	Jumlah Klaster
1	20	75
2	30	267
3	40	409
4	50	459
5	60	483
6	70	506
7	80	526
8	90	540

Distribusi jumlah klaster yang dihasilkan dari percobaan yang dilakukan dapat dilihat pada Gambar 5, bahwa semakin besar nilai batas threshold, maka jumlah klaster akan semakin besar.



Gambar 5 Hasil jumlah klaster

5. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan dapat disimpulkan hal-hal sebagai berikut :

1. Sistem dapat menampilkan dokumen yang mempunyai kesamaan/kemiripan Judul dan Abstrak dengan cara memasukkan teks yang diinginkan pengguna.
2. Penggunaan basisdata untuk menyimpan data indek dapat mempercepat proses pengindekan dokumen dan pengukuran kemiripan.
3. Hasil pengukuran kemiripan dengan cosine similarity untuk dokumen abstrak tersebut setelah dibandingkan antara pengukuran manual dan sistem, mempunyai tingkat akuransi 99%.
4. Klaster dapat membantu menemukan dokumen yang ada dalam satu klaster dengan query yang dimasukkan oleh pengguna.

Daftar Pustaka

Alex Nikolai, Barbara Hammer, Frank Klawonn, 2007. Single pass clustering for large data sets, Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM, 2007), 1-6.

Baeza, R dan Ribeiro, B, 1998, Modern Information Retrieval, ACM Press New York USA

Klampanos Iraklis A., Joemon M. Jose, C. J. Keith van Rijsbergen, 2006. Single-Pass Clustering for Peer-to-Peer Information Retrieval: The Effect of Document Ordering, Proceedings of the First International Conference on Scalable Information Systems, May 29-June 1 2006

Porter, M., (1980), An algorithm for suffix stripping, Program13(3), 130-137.

Pressman R, 1997, Software Engineering, Mc Graw Hill, USA.

Salton, G., 1989, Automatic Text Processing, The Transformation, Analysis, and Retrieval of Information by Computer, Addison – Wesly Publishing Company, Inc. All rights reserved.

Salton, G. and Buckley, 1988, Term Weigting Approaches in Automatic Text Retrieval, Department of Computer Science, Cornell University, Ithaca, NY 14853, USA.

Salton, G., 1971, Cluster Search Strategies and the Optimization of Retrieval Efectiveness, dalam G. Salton, ed. The SMART Retrieval System: Experiments in Automatic Document Processing. Englewood Cliffs: Prentice-Hall, 223-242

Sambasivam Samuel, Nick Theodosopoulos, 2006. Advanced Data Clustering Methods of Mining Web Documents, Issues in Informing Science and Information Technology, Volume 3, 565-579

Schleimer Saul, Daniel S. Wilkerson, Alex Aiken, 2003, "Winnowing : Local Algorithms for Document Fingerprinting", SIGMOD 2003,June 9 - 12, 1-10

Sihombing P; Embong A dan Sumari P, 2005, Application of Genetic Algorithm to Determine A Document Similarity Level in IRS, The First Malaysian Software Engineering Conference, 167-171

Sihombing P; Embong A dan Sumari P, 2006, Comparison of Document Similarity in Information Retrieval System by Different Formulation, Proceedings of the 2nd IMT-GT Regional Conference on Mathematics, Statistics and Aplikation, Malaysia, June 13-15, 1-8

Sridevi. K, R. Umarani. V.Selvi, 2011, An Analysis of Web Document Clustering Algorithms, International Journal of Science and Technology 1 (6), 275-282.

Tala, Z, 2003, A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia, Master of Logic Project, Institute for Logic, Language and Computation, Universiteit van Amsterdam, The Netherlands.