ANALISA PERFORMA METODE COSINE DAN JACARD PADA PENGUJIAN KESAMAAN DOKUMEN

Sugiyamto [1], Bayu Surarso [2] dan Aris Sugiharto [3]

[1] Jurusan Sistem Informasi Universitas Stikubank
[2] Jurusan Matematika FSM Universitas Diponegoro
[3] Jurusan Ilmu Komputer/Informatika Universitas Diponegoro

Abstract

Saat ini teknologi informasi memudahkan distribusi data-data digital melalui berbagai media, salah satunya adalah dokumen. Namun hal ini menyebabkan adanya penyalahgunaan dalam bentuk duplikasi yang mengarah pada kegiatan plagiarism terutama untuk naskah-naskah akademik seperti skripsi atau tugas akhir. Berbagai metode dikembangkan untuk meminimumkan terjadinya duplikasi illegal. Salah satunya adalah dengan menggunakan teknik pengujian kemiripan. Pada penelitian ini dibandingkan performa dari metode Cosine dan Jaccard untuk menguji tingkat kemiripan dokumen dalam bentuk abstrak. Hasil penelitian menunjukkan bahwa pengujian kemiripan menggunakan Cosine memiliki tingkat akurasi lebih tinggi yaitu 0,949808 dibandingkan dengan Jaccard sebesar 0,949077.

Keywords : dokumen, kemiripan, cosine, jaccard.

1. PENDAHULUAN

Pada era teknologi informasi memungkinkan sebuah dokumen diubah menjadi bentuk digital. Begitu pula dengan dokumen akademik seperti skripsi, tesis maupun desertasi mahasiswa. Digitalisasi dipilih karena terdapat beberap hal yang menguntungkan diantaranya adalah murah, mudah digandakan atau diduplikasi tanpa terjadi penurunan kualitas. Pada sisi lain digitalisasi juga menyebabkan dokumen-dokumen disalin dan digunakan untuk aktifitas-aktifitas yang tidak pada tempatnya, seperti menggunakan naskah akademik orang lain dengan cara disalin atau copy paste menjadi sebuah naskah akademik yang seolah-olah asli tanpa menyebutkan sumbernya atau lebih deikenal dengan nama plagiarism.

Dengan maraknya kegiatan plagiarism maka penelitian yang difokuskan pada deteksi kesamaan dokumen menjadi sangat mendesak. Beberapa penelitian untuk mengetahui tingkat kesamaan dokumen telah dilakukan, diantaranya oleh Strasberg (2002) yang meneliti kemampuan pencarian kemiripan dokumen yang diimplementasikan pada sebuah artikel berita dan jurnal. Kemudian Saul Schleimer, (2003), mendeteksi satu sifat penting dokumen local dengan menggunakan algoritma sidik jari (*fingerprinting*) dokumen dengan aplikasi MOSS (*Measure Of Software Similarity*), yang, memberikan hasil eksperimen pada data Web dengan hasil kinerja kurang lebih 33%.

Selanjutnya Sihombing (2005), menggunakan algoritma genetika untuk menentukan tingkat kesamaan dokumen pada sistem temu kembali informasi. Pada penelitian ini terungkap bahwa semakin banyak dokumen yang digunakan sebagai sumber dari query, maka tingkat kesamaan dokumen dapat menurun.

Pada tahun 2006, Sihombing juga meneliti penerapan algoritma genetika dengan menggunakan tiga formulasi yaitu : *Jaccard Formulation, Dice Formulation* dan *Horng & Yeh Formulation*. Hasil penelitian menunjukkan bahwa tren keseluruhan teknik probabilitas dalam kesamaan dokumen dengan menggunakan formulasi yang berbeda, jika jumlah permintaan meningkat, persentase kesamaan dokumen diambil bisa meningkat atau menurun.

Penelitian untuk mendeteksi kemiripan dokumen dengan mengembangkan perangkat lunak sederhana juga dilakukan oleh Yaakov HK

(2010), dengan membangun corpus (C), berisi 10.100 makalah akademik dalam ilmu komputer ditulis dalam bahasa Inggris dan dua tes set termasuk surat-surat yang dipilih secara acak dari C. Metode yang digunakan pada penelitian ini adalah *Full fingerprint methods* (FF), dimana menunjukkan hasil 53,4 % positif benar (*true positives*).

Adapun Tan et. All (2005) menggunakan *Jaccard similarity* atau *Jaccard Coefficient* untuk menghitung similarity antara dua *objects* (items). Seperti halnya *cosine distance* dan *matching coefficient* yang didasarkan pada *vector space similarity measure*.

Pada penelitian ini digunakan algoritma *single pass clusterring* untuk mengklasifikasikan dokumen dan mengujinya dengan pendekatan kesamaan dokumen yang berbeda yaitu menggunakan metode *cosine similarity* dan *Jaccard similarity*.

2. TINJAUAN PUSTAKA

Sistem Temu Kembali Informasi

Sistem Temu Kembali Informasi merupakan suatu sistem 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).

Sistem temu kembali teks (text retrieval) adalah sistem penemuan kembali informasi dalam bentuk dokumen dengan mengukur kemiripan (similarity) antara informasi yang tersimpan dalam basis data dengan query yang dimasukkan oleh pengguna (Baesa dan Ribeiro, 1998).

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.

Klastering Dokumen

Klastering biasanya 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.

Secara umum klastering dokumen adalah proses mengelompokkan dokumen berdasarkan kemiripan antara satu dengan yang lain dalam satu klaster (Gordon, 1991; Ellis, 1996).

Tujuan klastering dokumen adalah untuk memisahkan dokumen yang relevan dari dokumen yang tidak relevan. Atau dengan kata lain, dokumen-dokumen yang relevan dengan suatu query cenderung memiliki kemiripan satu sama lain dari pada dokumen yang tidak relevan, sehingga dapat dikelompokkan ke dalam suatu klaster. Klastering dokumen dapat dilakukan sebelum atau sesudah proses temu kembali. Pada klastering dokumen yang dilakukan sebelum proses temu kembali informasi, koleksi dokumen dikelompokkan ke dalam klaster berdasarkan kemiripan (*similarity*) antar dokumen. Selanjutnya dalam proses temu kembali informasi, apabila suatu dokumen ditemukan maka seluruh dokumen yang berada dalam klaster yang sama dengan dokumen tersebut juga dapat ditemukan. (Jian Zhang, dkk., 2001)

Dalam Sistem Temu Kembali Informasi, klastering dokumen memberikan beberapa manfaat, antara lain:

1. Mempercepat pemrosesan query dengan menelusur hanya pada sejumlah kecil anggota atau wakil klaster, sehingga dapat mempercepat proses temu kembali informasi
2. Membantu melokalisasi dokumen yang relevan
3. Membentuk kelas-kelas dokumen sehingga mempermudah penjelajahan dan pemberian interpretasi terhadap hasil penelusuran

4. Meningkatkan efektivitas dan efisiensi temu kembali informasi dan memberikan alternatif metode penelusuran

Selain itu, penggabungan antara penelusuran secara menyeluruh (full search) dengan penelusuran berbasis klaster (*cluster-based retrieval*) dapat meningkatkan ketelitian sampai dengan 25%. Hal yang sama dikemukakan oleh (Jian Zhang, dkk., 2001) bahwa penggabungan antara metode pengklasteran dengan fusion (pemberian peringkat terdadap dokumen secara keseluruhan) akan meningkatkan efektivitas temu kembali informasi. Pada algoritma klastering, dokumen akan dikelompokkan menjadi klaster-klaster berdasarkan kemiripan satu data dengan yang lain. Prinsip dari klastering adalah memaksimalkan kesamaan antar anggota satu klaster dan meminimumkan kesamaan antar anggota klaster yang berbeda.

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. Imbuhan (affixes) pada Bahasa Indonesia lebih kompleks bila dibandingkan dengan imbuhan (affixes) pada Bahasa Inggris. Hal ini disebabkan imbuhan (affixes) pada Bahasa Indonesia terdiri dari awalan (prefixes), sisipan (infixes), akhiran (suffixes), bentuk perulangan (repeated forms) dan confixes (kombinasi dari awalan dan akhiran). Imbuhan-imbuhan yang melekat pada suatu kata harus dihilangkan untuk mengubah bentuk kata tersebut menjadi bentuk kata dasarnya.

Stemming teks berbahasa Indonesia memiliki beberapa masalah yang sangat khusus terhadap bahasa. Salah satu masalah tersebut adalah perbedaan tipe dari imbuhan-imbuhan (affixes), yang lain adalah bahwa awalan (prefixes) dapat berubah tergantung dari huruf pertama pada kata dasar. Sebagai contoh ”me-” dapat berubah menjadi ”mem-” ketika huruf pertama dari kata dasar tersebut adalah ”b”, misalnya ”membuat” (to make), tetapi ”me-” juga dapat berubah menjadi ”meny-” ketika huruf pertama dari kata dasar melekat adalah ”s”, misalnya ”menyapu” (to sweep).

Selanjutnya ketika ada lebih dari satu imbuhan (affixes) yang melekat pada suatu kata, maka urutan untuk menghilangkan imbuhan-imbuhan (affixes) pada kata tersebut menjadi sangat penting. Jika dalam proses meghilangkan imbuhan-imbuhan (affixes) tersebut kita tidak memperhatikan urutan penghilangan imbuhan-imbuhan (affixes) tersebut, maka kata dasar yang benar dari kata tersebut tidak akan ditemukan. Sebagai contoh pada kata ”di-beri-kan” (to be given) yang diturunkan dari kata dasar”beri” (to give). Jika kita menghilangkan akhiran (suffixes) ”kan” terlebih dahulu sebelum menghilangkan awalan (prefix) ”di-” maka pada proses stemming ini kita mendapatkan kata dasar yang benar yaitu ”beri” (to give), akan tetapi jika algoritma stemming mencoba untuk menghilangkan awalan (prefixes) terlebih dahulu sebelum akhiran (suffixes) maka hasil kata dasar yang dihasilkan dari proses stemming dengan menggunakan algoritma tersebut adalah ”ikan” (fish) (setelah menghilangkan awalan ”di” dan ”ber”) dimana ”ikan” merupakan kata dasaryang valid yang terdapat dalam kamus tetapi ”ikan” bukan merupakan kata dasar yang benar untuk kata turunan ”diberikan”.

Penelitian terhadap stemming untuk text retrieval, machine translation, document summarization dan text classification sudah pernah dilakukan sebelumnya. Untuk stemming yang dilakukan pada Text Retrieval, stemming ini meningkatkan kesensitivan retrival dengan meningkatkan kemampuan untuk menemukan document yang relevan, tetapi hal itu terkait dengan pengurangan pada pemilihan dimana pengelompokkan menjadi kata dasar menyebabkan penghilangan makna kata. Pada Text Retrival stemming diharapkan dapat meningkatkan recall, tetapi memungkinkan untuk menurunkan precision.

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, untuk masing-masing data d :

(1) loop

a) menemukan a klaster c yang memaksimalkanan fungsi objektif
b) jika nilai dari fungsi subjektif > a maka nilai ambang masukkan d di dalam 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).

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 (Salton, 1989). Untuk notasi himpunan dapat digunakan rumus

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

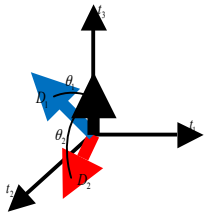
dimana $|X \cap Y|$ adalah jumlah term yang ada pada dokumen X dan yang ada pada dokumen Y, $|X|$ adalah jumlah term yang ada pada dokumen X dan $|Y|$ adalah umlah term yang ada pada dokumen Y .

Dari notasi himpunan di atas dapat dibuat persamaan matematika sebagai berikut :

$$Similarity(X,Y) = \frac{\sum_{i=1}^{i} x_i y_i}{\sqrt{\sum_{i=1}^{i} X_i^2 . \sum_{i=1}^{i} Y_i^2}}$$

dimana x dan y adalah dokumen yang berbeda, x_i adalah term i yang ada di dokumen x y_i adalah term i yang ada di dokumen y

Sedangkan pada proses nilai cosinus berdasarkan kata kunci pada setiap dokumen dapat diperlihatkan pada gambar 1 dimana D_1 adalah dokumen 1, D_2 adalah dokumen 2, Q adalah kata kunci dan t adalah kata dalam basisdata.



Gambar 1. Vektor similaritas kata kunci dengan dokumen (Salton, 1989)

Pada proses pencarian kemiripan antar dokumen yang bukan berdasarkan kata kunci maka variable Q diganti dengan kumpulan kata pada suatu dokumen yang akan dicari dokumen lain yang mirip dengan dokumen tersebut. Tidak ada proses query yang digenerate langsung oleh pengguna. Query dihasilkan dari kata-kata yang muncul di dokumen yang terpilih.

Pada proses pencarian kemiripan antar dokumen yang bukan berdasarkan kata kunci maka variable Q diganti dengan kumpulan kata pada suatu dokumen yang akan dicari dokumen lain yang mirip dengan dokumen tersebut. Tidak ada proses query yang digenerate langsung oleh pengguna. Query dihasilkan dari kata-kata yang muncul di dokumen yang terpilih.

Jaccard Coefficient

Jaccard Coefficient merupakan metode yang digunakan untuk menghitug tingkat kesamaan (similarity) antar dua buah objek. Untuk tujuan

klastering dokumen, fungsi yang baik adalah fungsi Jaccard Coefficient (Salton, 1989).
Untuk notasi himpunan dapat digunakan rumus

$$Similarity(X,Y) = \frac{|X \cap Y|}{|X| + .|Y| - |X \cap Y|}$$

dari notasi himpunan di atas dapat dibuat persamaan matematika sebagai berikut :

$$Similarity(X,Y) = \frac{\sum_{i=1}^{i} x_i y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2 - \sum_{i=1}^{t} x_i y_i}$$

Untuk merealisasi konsep ini, setiap dokumen harus dikorelasikan dengan subyek dengan relasi *many to many*, artinya satu subyek bisa memiliki beberapa dokumen, sebaliknya satu dokumen bisa juga memiliki beberapa subyek. Untuk dapat melakukan pengelompokan dokumen terhadap subyek dapat dilakukan dengan 2 cara, yaitu:

1. Memasukkan setiap dokumen secara langsung kedalam subyek
2. Memasukkan dokumen secara tidak langsung kedalam suatu subyek dengan menggunakan bantuan term.

Untuk dokumen dalam jumlah yang sangat banyak, tidak dilakukan pengelompokan dengan cara memasukkan satu persatu dokumen kedalam subyek, yaitu dengan memperhitungkan frekuensi kemunculan *term* dalam dokumen tersebut dan jumlah dokumen yang mengandung *term* tersebut.

3. METODOLOGI PENELITIAN

Alat dan Bahan

Pada penelitian ini digunakan data sejumlah 550 skripsi mahasiswa dalam bentuk judul dan abstrak. Sedangkan spesifikasi alat meliputi hardware dan software dengan spesifikasi sebagai berikut :

a. Perangkat keras
   Spesifikasi perangkat keras yang digunakan *processor* Intel *Core* i5, 2.30 GHz, RAM 4 GB, Hardisk 360 GB, *monitor*.

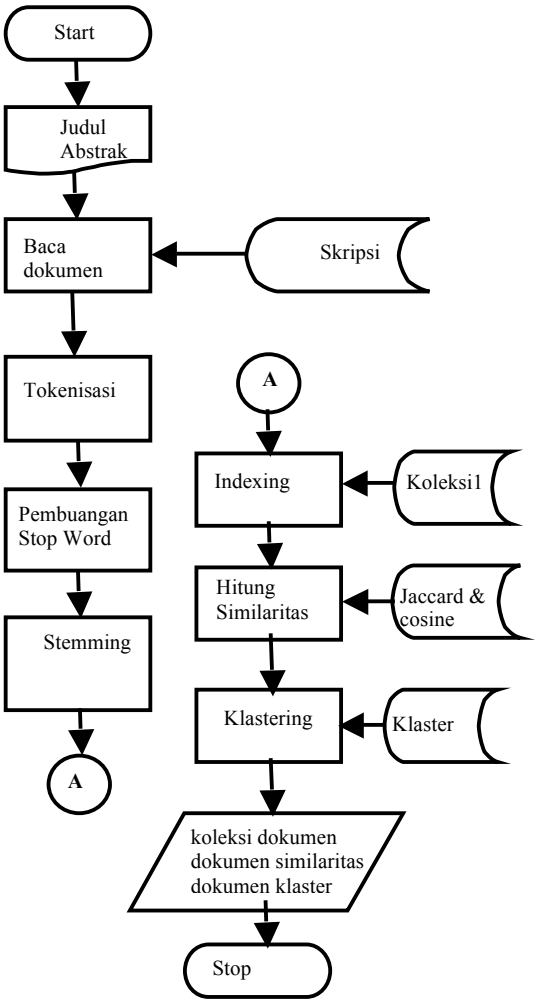
b. Perangkat lunak
   Perangkat lunak yang digunakan dalam sistem ini adalah Microsoft Windows 7 Ultimate, Bahasa pemrograman PHP, webserver Apache 2.0 dan server basis data MySql 5.

Tahapan Penelitian

Penelitian ini memiliki tahapan-tahapan sebagai berikut :

a. Tokenisasi
b. Pembuangan Stop Word
c. Steming
d. Indexing
e. Similarity
f. Clustering

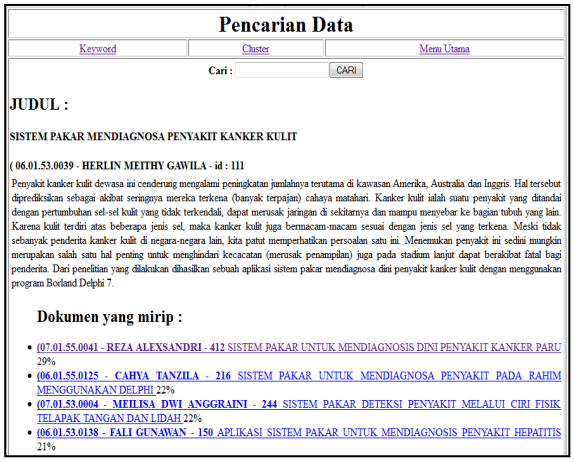
yang diperlihatkan pada gambar 2.



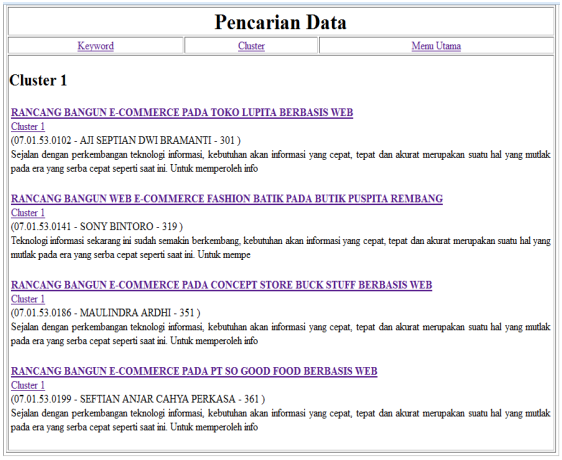
Gambar 2. Tahapan Penelitian

Hasil dan Pembahasan

Adapun antar muka yang digunakan untuk proses pencarian data sekaligus untuk melihat nilai kemiripannya dan proses klastering dapat dilihat pada gambar 3 dan 4.



Gambar 3. Antar muka pencarian data.



Gambar 4. Antar muka untuk pengklasteran

Setelah dilakukan serangkaian ujicoba dengan menggunakan perhitungan kemiripan Cosine similarity dan Jaccard seperti tahapan pada gambar 2, diperoleh hasil sebagai berikut : Untuk pengukuran kemiripan dengan menggunakan Cosine similarity diperlihatkan pada tabel 1 dan 2 sedangkan kemiripan dengan menggunakan Jaccard diperlihatkan pada tabel 3 dan 4.

Tabel 1. Kemiripan Judul dengan Cosine

id	*Id2*	*Sim*
1	2	0.10651036644905
1	3	0.27846110922372
1	4	0
1	5	0.09733604497504
….	….	….
….	….	….
1	548	0.19346019072098
1	549	0.23681337150673
1	550	0.21192962186799

Tabel 2. Kemiripan Abstrak dengan Cosine

id	*Id2*	*Sim*
1	2	0.15693353713962
1	3	0.06963972856599
1	4	0.06675547584228
1	5	0.16351551682676
….	….	….
1	548	0.13879565466240
1	549	0.13238875968437
1	550	0.16886428524790

Tabel 3. Kemiripan Judul dengan Jaccard

id	*Id2*	*Sim*
1	2	0.10540925533895
1	3	0.27735009811261
1	4	0
1	5	0.096225044864938
..	..	….
..	..	….
1	548	0.19245008972988
1	549	0.23570226039552
1	550	0.21081851067789

Tabel 4. Kemiripan Abstrak dengan Jaccard.

id	*Id2*	*Sim*
1	2	0.15592252703961
1	3	0.068629718555892
1	4	0.065644374841183
1	5	0.16250440681665
..	..	….
..	..	….
1	548	0.13779456365239
1	549	0.13128874968237
1	550	0.1687631851389

Tabel 5. Data 10 nilai kemiripan tertinggi untuk abstrak.

No.	*id*	*Id2*	*Jaccard*	*Cosine*
1	301	351	1	1
2	299	372	0.97301	0.973113
3	221	222	0.97183	0.972832
4	86	425	0.95	0.95
5	33	67	0.94021	0.941215
6	381	308	0.93795	0.938054
7	400	401	0.93159	0.932593
8	301	319	0.93026	0.932362
9	319	351	0.93026	0.931262
10	204	240	0.92565	0.926652

Dari data di atas dapat dilihat bahwa rata-rata kemiripan untuk abstrak dengan menggunakan Cosine sebesar 0,949808 sedangan jika digunakan Jaccard rata-ratanya sebesar 0,949077.

4. KESIMPULAN

Hasil dari penelitian ini menunjukkan bahwa penggunaan pengukuran kemiripan abstrak baik dengan Cosine maupun Jaccar secara rata-rata mengindikasikan bahwa keduanya memiliki performa yang tinggi, namun jika dibandingkan terlihat bahwa pengukuran dengan menggunakan Cosine similarity memiliki tingkat akurasi yang lebih baik yaitu sebesar 0,949808 sedangkan Jaccard sebesar 0,949077.

5. DAFTAR PUSTAKA

Alex N, Barbara H dan 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

Huang A., 2008. Similarity Measures for Text Document Clustering, *Proceedings of the New Zealand Computer Science Research Student Conference*, 2008, 49-56

Klampanos I A., Joemon M. J, dan 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.

Tan, P. N., M. Steinbach and V. Kumar, 2005. *Introduction to Data Mining*, Addison Wesley.