Pemilihan Korpus Statis Bersesuaian dengan *Cosine Similarity* dan Penggunaan IDF Global Pada Penambahan Dokumen Baru

Utomo Pujianto[1)] Arya Yudhi Wijaya[2)],
[1)] Jurusan Teknik Informatika Universitas Muhammadiyah Gresik
[2)] Jurusan Teknik Informatika Institut Teknologi Sepuluh Nopember
Email: utomo.pujianto@yahoo.co.id [2)], arya@if.its.ac.id [1)]

Abstrak – Permasalahan yang muncul pada saat pembobotan menggunakan nilai "term frequency–inverse document frequency" (tf-idf) adalah adanya kebutuhan untuk selalu melakukan perhitungan ulang nilai inverse document frequency (idf) setiap kali dokumen baru ditambahkan ke dalam database. Hal ini menyebabkan peningkatan kompleksitas komputasi menjadi $O(N^2)$. Untuk menangani masalah tersebut, dalam paper ini diusulkan sebuah metode yang menggunakan cosine similarity dan sejumlah korpus statis yang telah didefinisikan sebelumnya. Cosine similarity digunakan untuk menghitung kemiripan nilai term frequency (tf) dokumen baru dengan rerata nilai tf dari setiap korpus statis yang ada dalam database. Nilai idf dari korpus statis yang memiliki nilai similarity paling tinggi dengan dokumen baru kemudian dipilih sebagai nilai idf dari dokumen yang baru. Hasil uji coba menunjukkan bahwa tidak terdapat perbedaan yang signifikan antara nilai tf-idf yang dihitung dengan metode telah ada sebelumnya dengan metode yang diusulkan dalam paper ini. Dengan kata lain, metode ini dapat dipertimbangkan sebagai alternatif penentuan nilai idf, terutama karena kompleksitasnya yang hanya O(N).

Index Terms – cosine similarity, term weighting

I. Pendahuluan

Vektorisasi, atau pengubahan dokumen ke dalam Model Ruang Vektor (MRV) adalah langkah awal dalam Information Retrieval (IR). Pembobotan dilakukan dengan cara mencatat setiap term dalam dokumen dan menghitung frekuensinya sebagai bentuk representasi dokumen dalam MRV. Salah satu pilihan perhitungan pembobotan term dalam sebuah dokumen adalah *tf-idf* [1]. Penghitungan pembobotan dengan metode *tf-idf* terdiri dari komponen pembobotan lokal dan global. Pembobotan lokal dilakukan dengan menghitung frekuensi term yang muncul dalam sebuah dokumen (*tf*). Pembobotan global dilakukan dengan cara mencari di dokumen mana saja suatu term muncul dalam korpus (*idf*). Sebuah term dalam sebuah dokumen akan memiliki bobot yang tinggi apabila term tersebut memiliki frekuensi yang tinggi dan hanya muncul di sedikit dokumen dari seluruh dokumen yang ada.

Dalam sistem pembobotan dokumen menggunakan *tf-idf*, nilai *idf* akan selalu dihitung kembali setiap kali dilakukan penambahan dokumen baru ke dalam korpus. Dengan kata lain, penambahan dokumen dalam korpus akan menyebabkan *idf* seluruh dokumen akan berubah. Hal ini menimbulkan masalah, terutama pada kompleksitas komputasi yang mencapai $O(N^2)$.

Di lain pihak, Reed et al. [2] menemukan beberapa fakta penting tentang nilai frekuensi dokumen dari sebuah term, antara lain:

1. Nilai frekuensi dokumen dari sebuah term dalam sebuah korpus yang tidak diketahui ukurannya dapat diaproksimasi dengan menggunakan satu set dokumen yang sudah diketahui. Dengan kata lain, data set dalam ukuran yang lebih kecil kemungkinan besar dapat digunakan untuk mengaproksimasi data set yang lebih besar. Dengan catatan, ukuran korpus yang digunakan harus mencukupi untuk dapat mencakup kosa kata dalam jumlah besar.
2. Data frekuensi dokumen yang didapatkan dari satu sumber tertentu dapat secara efektif digunakan untuk mengestimasi sumber lain yang berbeda tetapi masih memiliki cukup kemiripan. Tetapi tidak cukup efektif untuk mengaproksimasi dokumen yang didapatkan dari sumber lain yang sama sekali berbeda.
3. Bahwa pada satu saat tertentu dari ukuran jumlah dokumen dalam korpus tidak terjadi peningkatan jumlah term yang unik (kosa kata) secara signifikan walaupun dilakukan penambahan sejumlah data baru pada korpus yang sudah ada.

Berdasarkan fakta-fakta tersebut dapat disimpulkan bahwa selama isi dokumen yang ditambahkan memiliki jenis topik yang mirip dengan korpus, perubahan nilai *idf* secara signifikan tidak akan berpengaruh pada perubahan nilai *tf-idf* [1], [2], [3].

Pendekatan *tf-icf* [2] menawarkan solusi dengan membuat korpus statis -- korpus yang tidak dapat ditambahi dokumen baru -- dengan banyak dokumen tertentu sebagai acuan pengambilan nilai *idf* sehingga apabila dokumen baru ditambahkan maka nilai *idf* diambilkan dari korpus statis tersebut. Akan tetapi, sebelumnya dilakukan pembatasan bahwa dokumen yang baru haruslah dokumen yang memiliki topik sejenis dengan topik yang dimiliki oleh korpus statis

yang telah tersedia. Hal ini juga menjadi masalah karena masih diperlukan langkah tambahan untuk menentukan apakah dokumen baru memiliki topik sejenis dengan korpus statis yang ada.

Paper ini mengusulkan sebuah metode baru untuk memperoleh nilai *idf* dari sebuah dokumen baru yang akan dimasukkan ke dalam database tanpa harus melakukan perhitungan ulang terhadap nilai *idf* untuk keseluruhan korpus yang ada di dalam database. Metode yang diusulkan merupakan perbaikan dari metode yang diusulkan dalam [2], yaitu dengan menambahkan otomisasi menggunakan *cosine similarity* dari nilai *tf* dokumen baru yang dibandingkan terhadap nilai *tf* korpus statis.

II. Metode yang diusulkan

Dalam paper ini diusulkan sebuah metode penentuan nilai *idf* dokumen baru yang akan ditambahkan, sebelum dilakukan penghitungan *tf-idf* sehingga dokumen baru yang dimasukkan akan memiliki nilai *idf* yang lebih tepat. Secara garis besar disediakan beberapa korpus statis yang memiliki jenis topik berbeda-beda dan memiliki nilai *idf* yang berbeda pula. Ketika dokumen baru ditambahkan, maka nilai *idf* yang dipakai adalah nilai *idf* yang memiliki jenis topik yang sesuai dengan dokumen baru yang ditambahkan. Pengenalan jenis topik tersebut dilakukan dengan memanfaatkan nilai *tf* dari dokumen baru.

Frekuensi term $tf_{t,d}$ dari term *t* dalam dokumen *d* didefinisikan sebagai banyaknya kemunculan *t* dalam dokumen *d*. Dikarenakan nilai $tf_{t,d}$ dalam suatu dokumen *d* sangat bervariasi dan memiliki jangkauan yang sangat besar, maka dilakukan normalisasi *log* pada $tf_{t,d}$ sehingga

$$w_{t,d} = \begin{cases} 1 + \log 10\ \mathrm{tf}_{t,d}, & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0, & \text{otherwise} \end{cases} \qquad (1)$$

Penghitungan $tf_{t,d}$ masih menimbulkan kekurangan dikarenakan semakin besar $tf_{t,d}$ tidak menjamin tingkat kesesuaian query dengan dokumen. Misalnya pada jenis kata hubung yang akan selalu memiliki frekuensi yang besar walaupun kata hubung bukanlah term yang signifikan yang mewakili suatu dokumen.

Oleh karena itu diperlukan suatu besaran yang berfungsi sebagai pengukur tingkat signifikansi suatu term dalam suatu dokumen. Secara alamiah, semakin sedikit term tersebut muncul di seluruh dokumen maka semakin signifikan term tersebut untuk mewakili sebuah dokumen. Misalnya kata hubung, kata hubung hampir muncul pada seluruh dokumen yang ada sehingga kata hubung kurang signifikan digunakan sebagai wakil dari dokumen. Besaran yang berfungsi sebagai pengukur tingkat signifikansi suatu term dalam suatu dokumen ini disebut dengan inverse document frequency idf_t yang didefinisikan sebagai

$$idf_t = \frac{\log_{10} N}{df_t} \qquad (2)$$

dimana *N* adalah banyaknya dokumen dalam koleksi, sedangkan df_t adalah ada di berapa dokumen term *t* dapat ditemukan dalam *N* dokumen.

Penambahan dokumen baru pada korpus akan berdampak pada perubahan nilai *idf* seluruh dokumen dalam korpus sehingga memakan ongkos komputasi sebesar $O(N^2)$. Oleh karena itu disusulkan metode untuk memangkas ongkos komputasi menjadi $O(N)$ dengan cara membuat korpus statis yang mewakili nilai *idf* [2].

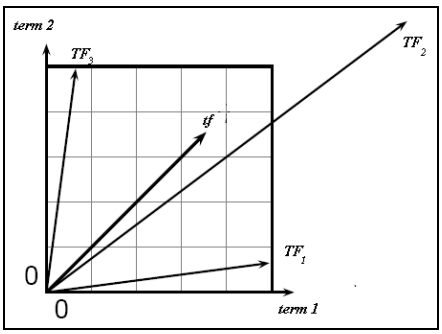
Agar dokumen baru yang ditambahkan dapat diberikan nilai *idf* yang tepat, maka disediakan beberapa korpus statis yang memiliki memiliki TF_{ij} dan idf_i dimana TF_{ij} adalah term frekuensi korpus *i* dokumen *j* dan idf_i adalah *idf* korpus *i*.

$$TF_{ij} = \sum_j tf \qquad (3)$$

Apabila frekuensi term dokumen baru pada *tf* maka kemiripan dokumen baru dengan korpus-korpus statis yang ada dapat dihitung dengan menggunakan *cosine similarity* yaitu:

$$\cos(\vec{tf}, TF_i) = \frac{\vec{tf} \bullet TF_i}{|\vec{tf}||\vec{d}|} = \frac{\vec{tf}}{|\vec{tf}|} \bullet \frac{TF_i}{|TF_i|} = \frac{\sum_{j=1}^{|V|} tf_j TF_{ij}}{\sqrt{\sum_{j=1}^{|V|} tf_j^{\,2}} \sqrt{\sum_{i=1}^{|V|} TF_{ij}^{\,2}}}$$

Ilustrasi pengukuran tingkat kemiripan antara dokumen baru dengan korpus-korpus statis yang ada dapat dilihat di Gambar 1.

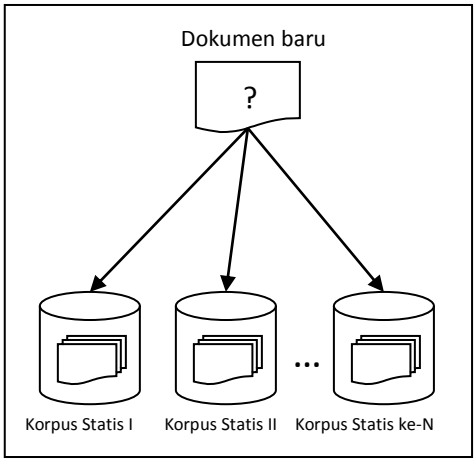


Gambar 1.
Ilustrasi MRV untuk menghitung cosine similarity dokumen baru dengan korpus statis

Metode yang diusulkan dapat didefinisikan sebagai serangkaian langkah sebagai berikut:

1. Membangun sejumlah korpus statis.
2. Menghitung nilai *tf* dan *idf* dari masing-masing korpus statis.
3. Menghitung nilai *tf* dari dokumen baru.
4. Membangun MRV berdasarkan nilai-nilai *tf* yang diperoleh, kemudian menghitung *cosine similarity* antara dokumen baru dengan korpus-korpus statis.

5. Melakukan perankingan terhadap nilai-nilai *cosine similarity* yang diperoleh untuk mendapatkan korpus yang paling mirip/sejenis dengan dokumen baru. Ilustrasi langkah 5 dapat dilihat pada Gambar 2.
6. Menetapkan nilai *idf* dokumen baru menjadi sama dengan nilai *idf* dari korpus statis yang paling mirip.



Gambar 2.
Membandingkan nilai *cosine similarity* vektor *tf* dokumen baru dan vektor *tf* korpus statis.

III. Uji Coba dan Pembahasan

Untuk mengevaluasi kinerja dari metode baru yang diusulkan, dilakukan serangkaian uji coba dengan menggunakan sejumlah dokumen yang dikumpulkan dari situs www.detiksport.com, www.detikhealth.com, dan www.detikfinance.com. Dari masing-masing situs web tersebut diambil sejumlah 125 dokumen secara acak untuk dijadikan sebagai korpus statis. Sehingga total dokumen yang dijadikan korpus statis adalah sebanyak 375 buah dokumen yang dibagi secara merata menjadi tiga macam korpus statis yaitu kesehatan, keuangan, dan sepak bola. Jumlah term unik untuk korpus statis kesehatan, keuangan, sepak bola berturut-turut adalah 3890, 3680 dan 3435.

Selain dokumen-dokumen yang akan digunakan sebagai korpus statis, dari masing-masing situs tersebut juga diambil 25 buah artikel untuk kemudian dijadikan sebagai data uji. Sehingga total dokumen baru yang akan dicari kemiripan topiknya dengan korpus statis yang paling sesuai adalah sejumlah 75 dokumen.

Dalam penelitian ini akan dilakukan dua macam uji coba. Uji coba pertama adalah uji coba untuk mengukur seberapa akurat *cosine similarity* dapat digunakan untuk menentukan korpus statis mana yang memiliki kemiripan topik dengan dokumen baru. Uji coba kedua dilakukan untuk mengetahui seberapa signifikan perbedaan nilai *tf-idf* metode yang diusulkan dibandingkan dengan *tf-idf* biasa yang selalu melakukan pembaruan nilai *idf* setiap dokumen dimasukkan.

Uji Coba Akurasi *Cosine Similarity* untuk Memilih Korpus yang Sesuai Dengan Dokumen Input

Uji coba dilakukan sebagaimana ilustrasi pada Gambar 2. Dokumen baru dimasukkan dan jenis dokumen baru tersebut belum diketahui apakah mirip dengan korpus satatis I atau korpus statis yang lain. Data uji terdiri dari 75 dokumen baru yang akan dikenali termasuk dalam kelas kesehatan, keuangan atau sepak bola dengan menggunakan *cosine similarity*. Dari 75 dokumen baru tersebut, sebelumnya diketahui bahwa 25 dokumen pertama adalah dokumen berita kesehatan, 25 selanjutnya adalah dokumen berita keuangan dan 25 yang terakhir adalah dokumen berita sepak bola.

Dari 25 dokumen kesehatan yang dimasukkan, 22 dokumen dikenali sebagai dokumen kesehatan, 2 dokumen masuk kelas keuangan dan 1 dokumen masuk kelas sepak bola. Dari 25 dokumen keuangan yang dimasukkan, 22 dokumen dikenali sebagai dokumen keuangan, 2 dokumen sebagai sepak bola dan 1 dokumen sebagai kelas kesehatan. Dari 25 dokumen sepak bola, 21 dokumen dikenali sebagai dokumen sepakbola, 2 dokumen kelas keuangan dan 2 dokumen masuk dalam kelas kesehatan. Tabel 1 menunjukkan hasil uji coba akurasi *cosine similarity* untuk memilih korpus yang sesuai dengan dokumen input sesuai dengan kondisi di atas.

Tabel 1.
Hasil ujicoba pemilihan corpus statis yang sesuai dengan dokumen baru.

Akurasi	Korpus Statis		
	Kesehatan	Keuangan	Sepakbola
Presisi	22/25	22/26	21/25
Recall	22/25	22/25	21/25

Uji Coba Perbedaan Signifikansi Hasil TF-IDF Metode yang Disulkan Dibanding TF-IDF Biasa

Uji coba kedua dilakukan untuk mengetahui seberapa signifikan perbedaan nilai *tf-idf* metode yang diusulkan dibandingkan dengan *tf-idf* biasa yang selalu melakukan pembaruan nilai *idf* setiap dokumen dimasukkan. Masing-masing korpus statis yang berisi dokumen kesehatan, keuangan dan sepakbola dimasukkan 5 dokumen baru yang akan dihitung bobotnya dengan metode *tf-idf*.

Tabel 2 menunjukkan hasil percobaan ketika dimasukkan 5 dokumen baru d_1, d_2, d_3, d_4 dan d_5 satu per satu secara berurutan yang diidentifikasi dengan benar oleh *cosine similarity* sebagai dokumen kesehatan. M_1 menunjukkan *idf* yang diusulkan yaitu dengan mengambil *idf* dari korpus statis kesehatan tanpa melakukan pembobotan ulang pada *tf-idf*. M_2 menunjukkan *tf-idf* sesuai dengan aturan normal, yaitu dilakukan pembaruan nilai *idf* setiap dimasukkan dokumen baru secara berurutan. Term yang ditampilkan

adalah t_{01}, $t_{02, \ldots,}$ t_{20} yang merupakan cuplikan term yang saling beririsan antara korpus kesehatan dengan dokumen d_1, d_2, d_3, d_4 dan d_5.

Tabel 3 menunjukkan hasil percobaan ketika dimasukkan 5 dokumen baru d_1, d_2, d_3, d_4 dan d_5 satu per satu secara berurutan yang diidentifikasi dengan benar oleh *cosine similarity* sebagai dokumen keuangan. M_1 menunjukkan *tf-idf* yang diusulkan yaitu dengan mengambil *idf* dari korpus statis kesehatan tanpa melakukan pembobotan ulang pada *idf*. M_2 menunjukkan *tf-idf* sesuai dengan aturan normal, yaitu dilakukan pembaruan nilai *idf* setiap dimasukkan dokumen baru secara berurutan. Term yang ditampilkan adalah t_{01}, $t_{02, \ldots,}$ t_{10} yang merupakan cuplikan term yang saling beririsan antara korpus kesehatan dengan dokumen d_1, d_2, d_3, d_4 dan d_5.

Tabel 4 menunjukkan hasil percobaan ketika dimasukkan 5 dokumen baru d_1, d_2, d_3, d_4 dan d_5 satu per satu secara berurutan yang diidentifikasi dengan benar oleh *cosine similarity* sebagai dokumen sepak bola. M_1 menunjukkan *tf-idf* yang diusulkan yaitu dengan mengambil *idf* dari korpus statis kesehatan tanpa melakukan pembobotan ulang pada *idf*. M_2 menunjukkan *tf-idf* sesuai dengan aturan normal, yaitu dilakukan pembaruan nilai *idf* setiap dimasukkan dokumen baru secara berurutan. Term yang ditampilkan adalah t_{01}, $t_{02, \ldots,}$ t_{20} yang merupakan cuplikan term yang saling beririsan antara korpus kesehatan dengan dokumen d_1, d_2, d_3, d_4 dan d_5.

Tabel 2
nilai bobot *tf-idf* metode yang diusulkan (M_1) dan metode standar (M_2) pada pemasukkan dokumen baru korpus kesehatan

Term	d_1			d_2			d_3			d_4			d_5		
	M_1	M_2	*Err*	M_1	M_2	*Err*	M_1	M_2	*Err*	M_1	M_2	*Err*	M_1	M_2	*Err*
t_{01}	3.97	3.81	0.16	4.97	4.68	0.29	36.19	36.71	0.52	11.39	10.59	0.8	5.59	5.53	0.07
t_{02}	4.16	3.98	0.18	9.39	9.15	0.24	6.97	6.02	0.94	19.86	18.67	1.2	6.53	6.41	0.12
t_{03}	4.38	4.17	0.21	6.53	6.39	0.15	4.16	4.02	0.14	16.63	15.95	0.68	1.72	1.69	0.03
t_{04}	4.38	4.17	0.21	4.97	4.68	0.29	2.21	2.12	0.1	9.29	8.81	0.48	28.04	28.04	0
t_{05}	4.97	4.66	0.31	4.64	4.42	0.23	3.06	3.02	0.04	8.64	8.22	0.41	2.97	2.92	0.04
t_{06}	4.97	4.66	0.31	4.64	4.42	0.23	1.11	1.14	0.03	4.97	4.67	0.3	1.88	1.84	0.04
t_{07}	4.97	4.66	0.31	4.64	4.42	0.23	2.97	2.93	0.03	4.64	4.4	0.24	1.72	1.69	0.03
t_{08}	8.32	7.95	0.36	4.97	4.68	0.29	4.97	4.7	0.27	6.97	5.99	0.98	4.52	4.46	0.06
t_{09}	8.84	8.68	0.16	7.59	7.36	0.24	5.97	5.44	0.53	6.97	5.99	0.98	3.38	3.31	0.07
t_{10}	129.14	119.46	9.69	34.74	34.02	0.72	19.86	18.8	1.06	62.28	60.93	1.36	18.75	18.94	0.19

Tabel 3
nilai bobot *tf-idf* metode yang diusulkan (M_1) dan metode standar (M_2) pada pemasukkan dokumen baru korpus keuangan

Term	d_1			d_2			d_3			d_4			d_5		
	M_1	M_2	*Err*	M_1	M_2	*Err*	M_1	M_2	*Err*	M_1	M_2	*Err*	M_1	M_2	*Err*
t_{01}	3.16	2.99	0.17	2.97	2.93	0.03	2.72	2.69	0.03	4.16	4.02	0.14	3.38	3.33	0.05
t_{02}	5.97	5.4	0.56	3.16	3.09	0.07	1.84	1.76	0.07	4.38	4.22	0.17	2.16	2.08	0.08
t_{03}	6.99	6.99	0	3.38	3.3	0.08	7.18	7.21	0.02	3.8	3.7	0.1	5.44	5.42	0.01
t_{04}	8.15	8	0.15	3.64	3.54	0.1	16.86	16.83	0.03	3.06	2.93	0.12	2.16	2.13	0.03
t_{05}	10.76	9.98	0.78	5.38	5	0.38	7.93	7.86	0.07	4.38	4.22	0.17	3.8	3.71	0.08
t_{06}	10.76	9.33	1.43	35.79	32.49	3.3	4.16	4.01	0.15	6.97	6.02	0.94	6.97	6.03	0.93
t_{07}	11.93	10.81	1.12	13.93	12	1.93	3.06	2.92	0.14	23.79	23.11	0.68	5.38	5.03	0.35
t_{08}	14.9	14	0.9	17.52	16.77	0.75	3.64	3.55	0.09	13.93	13.31	0.62	4.38	4.23	0.15
t_{09}	18.95	18.49	0.46	4.42	4.39	0.04	3.59	3.53	0.07	3.27	3.12	0.15	2.97	2.9	0.06
t_{10}	27.95	27.95	0	6.18	6.14	0.04	3.15	3.17	0.02	2.51	2.5	0.01	2.57	2.57	0

Tabel 4
nilai bobot *tf-idf* metode yang diusulkan (M_1) dan metode standar (M_2) pada pemasukkan dokumen baru korpus sepakbola

Term	d_1			d_2			d_3			d_4			d_5		
	M_1	M_2	*Err*	M_1	M_2	*Err*	M_1	M_2	*Err*	M_1	M_2	*Err*	M_1	M_2	*Err*
t_{01}	0.85	0.85	0.01	1.98	1.95	0.02	1.01	1.01	0.00	1.51	1.51	0.00	1.01	1.01	0.00
t_{02}	2.06	2.02	0.04	2.57	2.52	0.06	2.06	2.03	0.02	2.01	2.00	0.01	2.32	2.31	0.01
t_{03}	3.88	3.83	0.05	4.76	4.67	0.09	2.22	2.16	0.05	1.44	1.42	0.03	6.80	6.77	0.03
t_{04}	2.57	2.52	0.06	3.64	3.52	0.13	2.44	2.34	0.10	4.88	4.83	0.05	4.23	4.28	0.05
t_{05}	3.93	3.87	0.07	11.02	10.84	0.17	3.80	3.67	0.13	3.38	3.30	0.08	3.16	3.10	0.05
t_{06}	3.16	3.07	0.09	6.76	6.55	0.21	4.64	4.40	0.24	2.88	2.75	0.13	5.93	5.85	0.08
t_{07}	3.97	3.81	0.16	4.97	4.66	0.31	4.97	4.67	0.30	4.64	4.42	0.23	3.97	3.84	0.12
t_{08}	10.03	9.81	0.21	8.76	8.34	0.42	5.97	5.40	0.56	4.97	4.68	0.29	5.76	5.52	0.24
t_{09}	7.59	7.31	0.28	5.97	5.39	0.57	5.97	5.40	0.56	8.76	8.39	0.38	8.32	8.02	0.29
t_{10}	5.97	5.39	0.57	6.97	5.98	0.99	6.97	5.99	0.98	9.93	9.36	0.58	5.97	5.43	0.54

IV. Kesimpulan

Metode penghitungan *tf-idf* dokumen baru yang dimasukkan ke korpus tanpa mengubah *idf* dari seluruh dokumen di korpus menghasilkan nilai *tf-idf* yang tidak memiliki perbedaan signifikan apabila dilakukan penghitungan *tf-idf* dengan cara biasa yaitu dengan melakukan pengubahan seluruh *idf* dari korpus yang ada. Akan tetapi, dokumen baru yang dimasukkan harus memiliki tipe yang mirip dengan korpus yang ada. Apabila dokumen yang dimasukkan memiliki tipe yang belum diketahui jenisnya, maka metode yang diusulkan yaitu dengan melakukan pre-proses pemilihan korpus mana dari *n* korpus yang tersedia yang paling sesuai dengan dokumen baru. Pemilihan korpus mana yang paling sesuai dengan dokumen baru dilakukan dengan efektif menggunakan *cosine similarity*.

Referensi

[1] Spärck Jones, Karen (1972). "A statistical interpretation of term specificity and its application in retrieval". Journal of Documentation 28 (1): 11–21

[2] Joel W. Reed, Yu Jiao, Thomas E. Potok, Brian A. Klump, Mark T. Elmore, and Ali R. Hurson. "*TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams, Machine Learning, and Applications*", ICMLA '06. 5th International Conference on Publication, pages: 258-263

[3] Viles, C.L., & French, J.C., "*On the update of term weights in dynamic information retrieval systems"*. In Proceedings of the Fourth International Conference on Information and Knowledge Management, pages 167-174.

[4] Viles, C.L., & French, J.C., "*Dissemination of collection wide information in a distributed information retrieval system"*. In Proceedings of SIGIR '95, PP. 12-20.

[5] Amit Singhal, "*Modern Information Retrieval: A Brief Overview*", In IEEE Data Engineering Bulletin 24(4), pages 35-43, 2001.

[6] G. Salton, A. Wong, and C. S. Yang (1975), "A Vector Space Model for Automatic Indexing," Communications of the ACM, vol. 18, nr. 11, pages 613–620.