SNATI
Seminar Nasional
Aplikasi Teknologi Informasi 2004
Yogyakarta, 19 Juni 2004

Sistem Pengidentifikasi Otomatis Keterkaitan Topik antar Paragraf dalam Dokumen Ekspositori

Rila Mandala, Andreas Prasetia, Rinaldi Munir, Harlili
Laboratorium Ilmu dan Rekayasa Komputasi, Departemen Teknik Informatika,
Institut Teknologi Bandung, Jalan Ganesha 10 Bandung, 40132.
e-mail: {rila, rinaldi, harlili)@if.itb.ac.id

Abstract

This paper conducted on identifying topical coherency between paragraphs of an expository text. Expository text is a kind of texts which is intended for publicity such as journal, news etc. Many expository texts consist of long sequences of paragraphs with very little structural demarcation, while others consist of sequences of paragraph which still discuss the same topic. This paper uses Similarity Coefficient to identify topical coherency between paragraphs. Similarity coefficient measures lexical similarity between adjacent paragraphs, with assumption that the more similar two adjacent paragraphs are, the more likely it is that the current topic continues. Pengidentifikasi Keterkaitan Topik Antar Paragraf (PKTAP) is a software for identifying topical coherency between paragraphs of an expository text, which is developed based on this method. Topical coherency identification made by PKTAP is compared against judgments made by human readers. The result shows that the readers have a high tendency to agree with PKTAP.

Keywords: *expository text, topical coherency, similarity coefficient*

1. Pendahuluan

Dokumen ekspositori adalah dokumen atau tulisan yang ditampilkan kepada publik seperti jurnal, berita atau artikel lainnya, dalam hal ini khususnya yang ditampilkan secara *online* melalui internet. Sekarang ini, dokumen ekspositori yang berukuran penuh atau panjang telah tersedia dalam jumlah besar, sedangkan abstraksi dan artikel pendek sudah dapat diakses bertahun-tahun yang lalu. Oleh karena itu, kebanyakan metode *information retrieval* yang ada lebih sesuai untuk mengakses abstraksi daripada dokumen-dokumen yang lebih panjang. Keberadaan dokumen berukuran penuh ini tentunya harus disertai dengan pendekatan baru dalam mengakses informasi [HEA93].

Suatu dokumen biasanya terdiri dari bermacam-macam topik, berbeda dengan abstraksi yang ringkas dan padat informasi. Identifikasi dan isolasi topik dengan membagi-bagi dokumen, yang disebut segmentasi teks, merupakan hal yang penting dalam pemrosesan bahasa alami, termasuk mesin penterjemah dan *information retrieval*. Dalam *information retrieval*, pengguna seringkali hanya tertarik pada topik (atau bagian) tertentu dari dokumen yang diambil, bukan pada keseluruhan dokumen itu. Untuk memenuhi kebutuhan tersebut, dokumen harus disegmentasi ke dalam bagian-bagian yang koheren atau berkaitan. Segmentasi dokumen ke dalam blok-blok teks dengan topik yang sama dapat membantu *search engine* untuk memilih dan mengambil suatu segmen yang sesuai dengan *query* yang diajukan pengguna. Segmentasi yang secara nyata dapat dilihat adalah adanya pembagian dokumen dalam paragraf-paragraf.

Dalam pendidikan formal diajarkan bahwa paragraf dituliskan sebagai suatu kesatuan yang utuh, memiliki kalimat pokok pikiran (pokok pembicaraan atau topik) dan dilengkapi dengan kalimat penjelasnya. Dalam kenyataannya, kondisi ini sering tidak terpenuhi. Penandaan paragraf tidak selalu digunakan untuk menunjukkan pergantian pokok pembicaraan, tetapi kadang hanya digunakan dalam tampilan fisik untuk membantu dalam pembacaan.

Struktur dari dokumen ekspositori dapat dikarakterisasi sebagai rangkaian topik atau pokok pembicaraan yang berhubungan dengan topik utamanya. Struktur topik seringkali ditandai dengan judul dan subjudul yang membagi dokumen ke dalam segmen yang berkaitan. Tetapi banyak juga dokumen ekspositori yang terdiri dari rangkaian paragraf yang panjang dengan batas-batas struktural yang tidak jelas, ataupun yang terdiri dari rangkaian paragraf yang masih membicarakan topik yang sama.

Makalah ini melaporkan penelitian yang dilakukan untuk :

a. Melakukan studi tentang metode yang dapat dipergunakan untuk mengidentifikasi keterkaitan topik antar paragraf dalam dokumen ekspositori.
b. Mempergunakan metode tersebut dalam mengidentifikasi keterkaitan topik antar paragraf dalam dokumen ekspositori.

2. Struktur Topik Dokumen Repositori

Struktur dari dokumen ekspositori dapat dikarakterisasi sebagai rangkaian topik atau pokok pembicaraan yang berhubungan dengan topik utamanya. Sebagai contoh adalah sebuah tulisan ilmiah yang populer berjudul *Stargazers*, dengan topik utamanya adalah keberadaan kehidupan di bumi dan planet-planet lainnya. Tulisan ini memiliki pokok-pokok pembicaraan sebagai berikut:

Paragraf	Topik atau Pokok Pembicaraan
1 – 3	*Intro – the search for life in space*
4 – 5	*The moon's chemical composition*
6 – 8	*How early proximity of the moon shaped it*
9 – 12	*How the moon helped life evolve on earth*
13	*Improbability of the earth-moon system*
14 – 16	*Binary/trinary star systems make life unlikely*
17 – 18	*The low probability of non-binary/trinary systems*
19 – 20	*Properties of our sun that facilitate life*
21	*Summary*

Struktur topik dari dokumen seringkali ditandai dengan judul dan subjudul yang membagi dokumen ke dalam segmen-segmen yang berkaitan. Tetapi banyak juga dokumen ekspositori yang terdiri dari rangkaian paragraf yang panjang dengan batas-batas struktural yang tidak jelas, ataupun yang terdiri dari rangkaian paragraf yang masih membicarakan topik yang sama.

Identifikasi struktur topik dari dokumen akan sangat berguna untuk aplikasi pemrosesan bahasa alami, seperti dalam peringkasan teks otomatis atau juga dalam

information retrieval. Algoritma *information retrieval* dapat menggunakan struktur topik untuk mengambil dan memberikan bagian-bagian yang penting dari suatu teks yang panjang. Salton dkk. (1993) telah melakukan penelitian menggunakan teks dari buku ensiklopedia, dan menyatakan bahwa *query* terhadap seksi dan paragraf memberikan hasil yang lebih baik dibandingkan dengan *query* terhadap keseluruhan dokumen [HEA94].

Morris dan Hirst (1991) mempelopori penelitian dalam komputasi struktur tulisan berdasarkan hubungan keterkaitan secara leksikal. Dengan menggunakan *thesaurus* yang lengkap (*Roget's Fourth Edition*), Morris telah mengembangkan suatu algoritma yang dapat menemukan rantai dari *term-term* yang berhubungan. Tetapi algoritma tersebut bertujuan untuk menemukan struktur *attentional/intentional*, berbeda dari yang dilakukan Hearst dalam *TextTiling*, yang menggunakan hubungan keterkaitan secara leksikal untuk membagi-bagi dokumen ekspositori ke dalam segmen-segmen yang mencerminkan struktur topiknya.

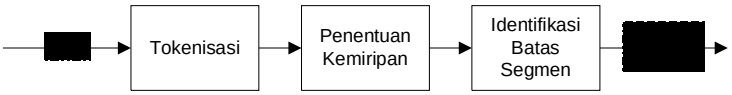
TextTiling mengasumsikan bahwa ada sekumpulan *item* leksikal yang digunakan dalam mendiskusikan suatu pokok pembicaraan, dan pada saat terjadi perubahan pokok pembicaraan, proporsi yang signifikan dari kata-kata yang digunakan juga berubah.

3. Algoritma Pencarian Struktur Topik

Peneliti-peneliti sebelumnya, seperti Halliday & Hasan (1976), Tannen (1989), Walker (1991) menyatakan bahwa pengulangan *term* menunjukkan indikasi yang kuat adanya keterkaitan. Hearst juga menemukan bahwa pengulangan *term* itu sendiri adalah indikasi yang sangat penting dalam struktur topik [HEA94].

Algoritma untuk pencarian struktur topik menggunakan pengulangan *term* sebagai indikasi keterkaitan secara leksikal adalah dengan membandingkan pasangan blok-blok teks berurutan, dan dihitung seberapa besar kemiripannya secara leksikal. Dalam metode ini diasumsikan bahwa semakin besar kemiripan antara dua blok teks, maka semakin besar kemungkinan bahwa topik itu berkelanjutan, dan sebaliknya jika semakin kecil kemiripannya, atau tidak mirip, maka hal itu menunjukkan adanya perubahan alur dari topik sebelumnya.

Algoritma utama pencarian struktur topik ini terdiri dari tiga bagian, seperti yang dapat dilihat pada Gambar 1, dan akan dijelaskan dalam sub-sub bab berikut.



Gambar 1. Algoritma utama pencarian struktur topik

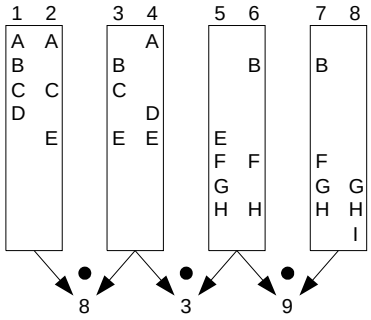
3.1 Tokenisasi

Tokenisasi adalah pembagian teks masukan menjadi kata-kata secara individu. Kemudian kata-kata tersebut dikelompokkan dalam blok-blok berukuran tertentu dan dinamakan *token-sequence*, sedangkan kata-kata yang umum dan tidak memiliki makna yang penting tidak diikutsertakan. Token atau *term* yang dianalisis dicatat dalam tabel, bersama dengan lokasi kemunculannya dalam *token-sequence* juga frekuensi atau banyaknya kemunculannya.

3.2 Penentuan Kemiripan

Setelah tokenisasi, langkah berikutnya adalah membandingkan kemiripan leksikal dari pasangan blok *token-sequence* yang berurutan.

Sebagai contoh, terdapat empat blok yang masing-masing terdiri dari dua *token-sequence*. Gambar 2 menunjukkan bagaimana skor kemiripan leksikal, yang disebut skor leksikal, dihitung dari pasangan blok yang dibandingkan.



Gambar 2. Perhitungan skor leksikal

Skor leksikal di sini adalah *inner product* dari dua vektor, di mana vektor tersebut tersusun atas frekuensi kemunculan token atau *term* dalam blok. Token A muncul satu kali dalam *token-sequence* pertama dan satu kali dalam *token-sequence* kedua, jadi token A muncul dua kali dalam blok pertama. Token A muncul kembali satu kali dalam blok kedua, yaitu pada *token-sequence* keempat. Perhitungan skor leksikal antara blok pertama dan kedua dapat dilihat dalam Tabel 1.

Tabel 1. Perhitungan skor leksikal

Token	***Frekuensi***		***Hasilkali***
	Blok 1	***Blok 2***	
A	2	1	2
B	1	1	1
C	2	1	2
D	1	1	1
E	1	2	2
		Jumlah	8

Dengan cara yang sama dihitung skor leksikal antara blok kedua dan ketiga, juga antara blok ketiga dan keempat.

Perhitungan skor leksikal di atas merupakan *inner product* yang tidak dinormalisasi. Normalisasi dilakukan dengan memberikan bobot *term* yang sesuai.

Term-term dalam dokumen mempunyai nilai penting *term* terhadap dokumen. Nilai penting *term*, secara kuantitatif dapat dihitung dengan beberapa faktor dalam dokumen. Penentuan nilai penting *term* ini disebut dengan pembobotan *term*.

Faktor pertama dalam penentuan bobot *term* adalah frekuensi *term* (*term frequency* atau *tf*). Setiap *term* yang mempunyai frekuensi tinggi dalam satu dokumen, berpengaruh terhadap pentingnya *term*, yang berarti meningkatkan bobotnya.

Faktor kedua adalah kejarangmunculan *term*. Setiap *term* yang tidak sering muncul di antara dokumen dalam koleksi dokumen, diberi bobot yang lebih tinggi daripada *term* yang sering muncul. Pembobotan akan memperhitungkan kebalikan frekuensi dokumen yang mengandung suatu *term*, yang kemudian disebut dengan *inverse document frequency* (*idf*) [GRO98].

Dalam *tf.idf* standar, yang digunakan dalam *information retrieval*, *term* yang sering muncul dalam sebuah dokumen tetapi relatif jarang muncul dalam keseluruhan koleksi dokumen dikatakan sebagai ciri utama dari dokumen tersebut.

Dalam segmentasi teks ini, setiap blok *token-sequence* dianggap sebagai satu bagian dari seluruh dokumen, dan frekuensi dari suatu *term* dalam masing-masing blok dibandingkan dengan frekuensi *term* tersebut dalam keseluruhan dokumen. Hal ini membantu menunjukkan perbedaan antara keberadaan *term* secara lokal dan global. Keberadaan *term* secara lokal adalah jika suatu *term* dibicarakan secara berulang-ulang dalam bagian yang terlokalisasi, dengan demikian menunjukkan bagian yang berkaitan. Sedangkan keberadaan *term* secara global adalah jika *term* tersebut sering muncul namun menyebar merata dalam keseluruhan dokumen.

Dengan demikian jika blok-blok yang berurutan memiliki banyak *term* yang sama, dan *term-term* tersebut memiliki bobot yang besar, maka hal ini menunjukkan bukti yang kuat bahwa blok-blok yang berurutan tersebut saling berkaitan.

Kemiripan antar blok paragraf dihitung dengan ukuran cosinus, yang disebut *Similarity Coefficient* (Koefisien Similaritas), sebagai berikut:

$$\cos(b_1, b_2) = \frac{\sum_{t=1}^{n} w_{t,b_1} w_{t,b_2}}{\sqrt{\sum_{t=1}^{n} w_{t,b_1}^2 \sum_{t=1}^{n} w_{t,b_2}^2}}$$

di mana *t* meliputi seluruh *term* dalam dokumen, w_{t,b_1} adalah bobot yang diberikan pada *term t* dalam blok b_1 dan w_{t,b_2} adalah bobot yang diberikan pada *term t* dalam blok b_2.

Jika koefisien similaritas antara dua blok nilainya tinggi, maka berarti bahwa kedua blok tersebut bukan hanya memiliki *term* yang sama, tetapi juga *term-term* yang sama itu secara relatif jarang muncul di bagian lain dalam dokumen. Tetapi tidak sebaliknya: jika blok yang berurutan memiliki ukuran kemiripan yang rendah, bukan berarti blok-blok tersebut tidak berkaitan.

Nilai-nilai koefisien similaritas yang diperoleh kemudian dapat digambarkan dalam grafik, dengan sumbu mendatarnya adalah nomor batas blok, yaitu batas antara dua buah blok *token-sequence*, dan sumbu tegaknya menunjukkan koefisien similaritas dari kedua blok tersebut.

Dari contoh pada Gambar 2 sebelumnya, ditabelkan frekuensi dari seluruh token seperti pada tabel 2.

Tabel 2. Frekuensi token dalam blok *token-sequence*

Token	***Frekuensi***			
	Blok 1	***Blok 2***	***Blok 3***	***Blok 4***
A	2	1	0	0
B	1	1	1	1
C	2	1	0	0
D	1	1	0	0
E	1	2	1	0
F	0	0	2	1
G	0	0	1	2
H	0	0	2	2
I	0	0	0	1

Dengan Persamaan 2.1 dihitung koefiesien similaritas antara blok *token-sequence* yang berurutan.

$$\cos(b_1, b_2) = \frac{(2)(1)+(1)(1)+(2)(1)+(1)(1)+(1)(2)}{\sqrt{(2^2+1^2+2^2+1^2+1^2)(1^2+1^2+1^2+1^2+2^2)}}$$

$$= \frac{8}{\sqrt{(11)(8)}}$$

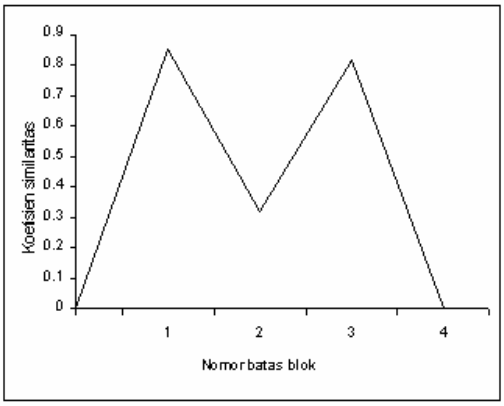
$$= 0.8528$$

Dengan cara yang sama diperoleh:

$\cos(b_2, b_3) = 0.3198$

$\cos(b_3, b_4) = 0.8182$

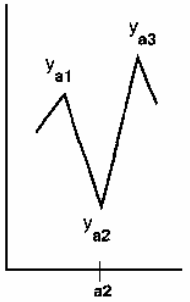
Dan grafiknya digambarkan seperti pada gambar 3.



Gambar 3. Plot koefisien similaritas

3.3 Identifikasi Batas Segmen

Batas segmen ditentukan dengan melihat perubahan dalam deretan koefisien similaritas. Untuk setiap batas blok, diperhatikan seberapa tajam kemiringan garis di kedua sisinya. Pertama, dilihat nilai-nilai di sebelah kiri suatu batas blok, terus bergerak selama nilainya menaik. Apabila sampai pada puncak, perbedaan koefisien antara puncak dan batas blok tersebut dicatat. Kemudian, dengan cara yang sama dilihat nilai-nilai di sebelah kanannya. Kedua buah nilai tersebut dijumlahkan, disebut nilai kedalaman, yaitu nilai yang menunjukkan seberapa tajam terjadi perubahan di kedua sisi batas blok token-sequence. Dengan Gambar 4 ditunjukkan bahwa nilai kedalaman untuk batas blok a2 adalah (ya1 - ya2) + (ya3 - ya2).



Gambar 4. Perhitungan nilai kedalaman

Nilai-nilai kedalaman yang didapatkan kemudian diurutkan, dan batas-batas blok dengan nilai kedalaman yang besar dipilih sebagai batas segmen, disesuaikan seperlunya mengikuti paragraf sesungguhnya.

Algoritma ini harus menentukan ke dalam berapa segmen suatu dokumen dibagi-bagi, mengingat bahwa setiap paragraf memiliki potensi untuk menjadi satu segmen. Berdasarkan rata-rata dan simpangan baku nilai-nilai kedalaman dari dokumen yang telah dianalisisnya, Hearst menyatakan bahwa suatu batas blok dapat dipilih sebagai batas segmen apabila nilai kedalamannya melebihi $\bar{x} - \sigma / 2$ [HEA94].

4. Evaluasi

Suatu cara untuk mengevaluasi algoritma segmentasi di atas adalah membandingkan segmentasi paragraph yang dihasilkan oleh sistem dengan segmentasi yang dilakukan oleh pembaca, atau dengan dokumen itu sendiri yang sudah dibagi-bagi oleh penulisnya.

Dalam pembandingannya dilakukan dua cara, yaitu:

a. Pembaca diminta untuk mengidentifikasi ada atau tidaknya keterkaitan topik antar paragraf dalam dokumen. Perangkat lunak PKTAP juga mengidentifikasi ada atau tidaknya keterkaitan topik antar paragraf dalam dokumen tersebut. Kemudian hasil identifikasi oleh pembaca dan oleh perangkat lunak PKTAP dibandingkan.
b. Dokumen yang digunakan: Autisme.
c. Perangkat lunak PKTAP mengidentifikasi ada atau tidaknya keterkaitan topik antar paragraf dalam dokumen, kemudian dokumen dengan hasil identifikasi tersebut ditunjukkan kepada pembaca. Pembaca diminta untuk menilai apakah kesimpulan yang diberikan oleh perangkat lunak PKTAP itu benar atau tidak, dengan menjawab SETUJU atau TIDAK SETUJU.

   Dokumen yang digunakan:
   - Galileo Meluncur, GPS (Bisa) Tergusur.
   - Menggugat Manajemen Keuangan Publik.
   - Soeharto Mampu Menjawab Sembilan Pertanyaan.

Hasil dari pengamatan disajikan dalam tabel 4.

Tabel 4. Ringkasan hasil pengamatan

Judul Dokumen	*Banyaknya Paragraf*	*Banyaknya Kata*	*Banyaknya Pembaca*	*Rata-rata Kesesuaian*
Autisme	12	1046	6	5.1818
Galileo Meluncur, GPS (Bisa) Tergusur	21	849	6	5.1000
Menggugat Manajemen Keuangan Publik	32	1410	6	5.3871
Soeharto Mampu Menjawab Sembilan Pertanyaan	20	777	5	4.2105

Rata-rata kesesuaian untuk dokumen pertama, “Autisme”, sebesar 5.1818, atau sekurang-kurangnya lima dari enam pembaca sependapat dengan perangkat lunak PKTAP.
Untuk dokumen kedua, “Galileo Meluncur, GPS (Bisa) Tergusur”, rata-rata kesesuaiannya sebesar 5.1000, atau sekurang-kurangnya lima dari enam pembaca sependapat dengan perangkat lunak PKTAP.

Untuk dokumen ketiga, “Menggugat Manajemen Keuangan Publik”, rata-rata kesesuaiannya sebesar 5.3871, atau sekurang-kurangnya lima dari enam pembaca sependapat dengan perangkat lunak PKTAP.

Untuk dokumen keempat, “Soeharto Mampu Menjawab Sembilan Pertanyaan”, rata-rata kesesuaiannya sebesar 4.2105, atau sekurang-kurangnya empat dari lima pembaca sependapat dengan perangkat lunak PKTAP.

Dapat dilihat bahwa pembaca yang melakukan identifikasi keterkaitan topik antar paragraf dalam dokumen yang diamati memiliki kecenderungan yang tinggi untuk setuju dengan identifikasi yang dilakukan oleh perangkat lunak PKTAP.

5. Kesimpulan

Kesimpulan yang dapat diambil dari pengembangan perangkat lunak PKTAP (Pengidentifikasi Keterkaitan Topik Antar Paragraf) adalah sebagai berikut:

a. Koefisien Similaritas dapat dipakai untuk mengukur seberapa besar kemiripan antara dua buah paragraf dalam dokumen ekspositori.
b. Perangkat lunak PKTAP yang telah dibangun dapat mengidentifikasi keterkaitan topik antar paragraf dalam dokumen ekspositori.

6. Ucapan Terima Kasih

Penelitian yang dilakukan oleh penulis di Laboratorium Ilmu dan Rekayasa Komputasi, Departemen Teknik Informatika ITB ini sebagian didanai oleh dana RUT (Riset Unggulan Terpadu) IX dari KMNRT (Kantor Menteri Negara Riset dan Teknologi), oleh karena itu penulis mengucapkan terima kasih.

Daftar Pustaka

[GRO98] Grossman, David A., and Frieder Ophir, *Information Retrieval: Algorithms and Heuristics*, Kluwer Academic Publisher, Massachusetts, 1998.

[HEA93] Hearst, Marti A., and Christian Plaunt, 1993, *Subtopic Structuring for Full-Length Document Access*, dalam Proceedings of SIGIR, Pittsburgh, 1993. http://www.sims.berkeley.edu/~hearst/publications.shtml

[HEA94] Hearst, Marti A., *Multi-Paragraph Segmentation of Expository Text*, dalam ACL, Las Cruces, 1994. http://www.sims.berkeley.edu/~hearst/publications.shtml

[HEA97] Hearst, Marti A., *TextTiling: Segmenting Text into Multi-paragraph subtopic passages*, Computational Linguistics, vol. 23 (1), 1997. http://acl.ldc.upenn.edu/J/J97/

[KAN98] Kan, Min-Yen, Judith L.Klavans and Kathleen R.McKeown, *Linear Segmentation and Segment Significance*, dalam Proceeding of the Sixth Workshop on Very Large Corpora, Canada, 1998.

[KOZ93] Kozima, Hideki, *Text segmentation based on similarity between words*, dalam Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Columbus, 1993.

[PRE01] Pressman, Roger S., *Software Engineering, A Practitioner’s Approach*, 5th ed., MacGraw-Hill, New York, 2001.