Seminar Nasional Aplikasi Teknologi Informasi 2005 (SNATI 2005)
Yogyakarta, 18 Juni 2005
ISBN: 979-756-061-6

PEMBENTUKAN INTISARI TOPIK SECARA OTOMATIS DALAM SUATU PARAGRAF DENGAN MODEL *VECTOR SPACE MODEL*

Muhammad Erwin Ashari Haryono
Laboratorium Pemrograman dan Informatika Teori, Jurusan Teknik Informatika, Fakultas Teknologi Industri, Universitas Islam Indonesia, Kampus Terpadu UII, Jl. Kaliurang Km 14.5 Yogyakarta
E-mail: meah@fti.uii.ac.id

Abstract

This paper is the implementation of both topic model text summarization and Information retrieval model. This paper conducted on identyfing topical coherency inside one paragraphs of an expository text. Expository text is a kind of text which is intended for publicity such as journal, news etc. Many expository text consist of long sequences of paragraphs with very little structural demarcation, while others consist of sequences of paragraphs which still discuss the same topic. This paper uses vector space model to identify topical coherency between paragraphs. Vector space model 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. We use expository text from one week daily local newspaper "Kedaulatan Rakyat" from cultural and educational topic. We choose 20 sample for testing our method, the result is not very good for the precision among main idea and its paragraphs (approximately 60 percentage success and 40 failured).

Keyword: *expository text, topical coherency, vector space model model*

1. Pendahuluan

Penelitian ini adalah studi kasus khusus dari bidang pengolahan bahasa alami (*natural language processing*) yang memfokuskan pada pencarian topik inti dari suatu paragraf. Penentuan topik secara alami oleh manusia dapat sangat mudah untuk dilakukan dikarenakan kemampuan manusia untuk dapat merangkum dan menentukan kata-kata mana yang bersesuaian yang dapat membentuk suatu kesatuan cerita, walaupun terkadang manusia pun akan menemukan kesulitan juga dalam penentuan itu. Dalam bidang komputer terutama Informatika, hal ini dibahas lebih lanjut dalam studi kepustakaan antara kemampuan tata bahasa linguistik, pembentukan *parsing*, aturan, dan pemrograman yang cukup efisien. Penentuan topik tersebut dilakukan dengan membagi (*segmentation*) kalimat dalam suatu alinea paragraf yang koheren dan berkaitan. Segmentasi paragraf ke dalam blok-blok teks dengan membantu mesin pencari untuk melakukan pembagian atau *clustering* terhadap topik-topik yang sama dalam beberapa alinea paragraf nantinya. Dalam pengolahan bahasa alami hal tersulit adalah bagaimana mengidentifikasi makna semantik dari suatu wacana. Tingkatan pendefinisian dan pengenalan makna secara semantik adalah tingkatan ke tiga tersulit setelah proses pengenalan makna morphologi (proses pembentukan kata), pola sintaksis (susunan antar bentuk kata dalam kalimat). Proses *parsing* juga akan sangat berpengaruh pada penentuan nilai semantik suatu kata yang dilihat dari makna yang ganda (ambiguinitas). Dalam penelitian ini tidak dikembangkan proses *parsing* yang dimaksud di atas.

2. Wacana Paragraf

Dalam pendidikan kebahasaan diajarkan bahwa paragraf dituliskan sebagai suatu kesatuan yang utuh, memiliki kalimat pokok pikiran (main topic, main discussion) dan dilengkapi dengan kalimat penjelasannya. Dalam kenyataannya, kondisi ini sering tidak terpenuhi. Penandaan paragraf tidak selalu digunakan dalam tampilan fisik untuk membantu dalam pembacaan. Struktur dari dokumen *ekspository* 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 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. Struktur topik dalam dokumen seringkali ditandai dengan judul dan subjudul yang membagi dokumen ke dalam segmen-segmen yang berkaitan. Tetapi banyak juga dokumen yang terdiri dari rangkaian paragraf yang panjang dengan batas-batas struktural yang tidak jelas, ataupun yang terdiri dari rangkaian-rangkaian yang masih membicarakan topik yang sama. [MAN04].

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 Hist (1991) mempelopori penelitian dalam komputasi struktur tulisan berdasarkan hubungan keterkaitan secara leksikal. Dengan menggunakan *thesaurus* yang lengkap (*Roget's Fourth Edition*), Morris telah

 ISBN: 979-756-061-6

mengembangkan suatu algoritma yang dapat menenemukan rantai dari *term-term* yang berhubungan. Tetapi algoritma tersebut bertujuan untuk menemukan struktur attentional/intentional, berbeda dari yang dilakukan Hears dalam *Texttilling*, yang menggunakan hubungan keterkaitan secara leksikal untuk membagi-bagi dokumen ke dalam segmen-segmen yang mencerminkan struktur topiknya. [MAN04]

3. Model Ruang Vektor

Banyak terdapat model dalam pencarian suatu dokumen informasi. Model tersebut antara lain adalah probabilistik, boolean (digolongkan ke dalam bentuk manual). Model Ruang vektor, latent semantic indexing (digolongkan ke dalam bentuk otomatis) dan model *fuzzy* dan genetika (digolongkan ke dalam bentuk adaptive). Pada penelitian ini digunakan model ruang vektor. Model ruang vektor pada umumnya dilakukan untuk melihat nilai kemiripan suatu dokumen terhadap suatu *query*. Pembobotan secara otomatis biasanya berdasarkan jumlah kemunculan suatu istilah dalam sebuah dokumen (*term frequency*/**tf**) dan jumlah kemunculannya dalam koleksi dokumen (*inverse document frequency*/**idf**). Bobot suatu istilah semakin besar jika istilah tersebut sering muncul dalam suatu dokumen dan semakin kecil jika istilah tersebut muncul dalam banyak dokumen. [GRO98].

Saat mesin menerima *query*, mesin akan membangun sebuah vektor Q (w_{q1},w_{q2},…w_{qt}) berdasarkan istilah-istilah pada *query* dan sebuah vektor D (d_{i1},d_{i2},…d_{it}) berukuran t untuk setiap dokumen. Pada umumnya SC dihitung dengan rumus *Cosine Measure* seperti persamaan 1 di bawah ini : [GRO98]. Selain persamaan 1 di bawah ini juga terdapat berbagai persamaan yang dapat dilakukan, persamaan tersebut adalah pers 6, pers 7 dan pers 8.

$$SC(Q, D_i) = \frac{\sum_{j=1}^{t} (w_{qj} * d_{ij})}{\sqrt{\sum_{j=1}^{t} (w_{qj})^2 \sum_{j=1}^{t} (d_{ij})^2}} \qquad (1)$$

dimana:

w_{qj} = bobot istilah j pada *query* q = $frek_{qj}$ *idf_j

d_{ij} = bobot istilah j pada dokumen i = tf_{ij} * idf_j

tf_{ij} = *term frequency* = kemunculan istilah t_j pada dokumen D_i

idf_j = *inverse document frequency* = $\log\left[\frac{d}{df_j}\right]$

d = jumlah total dokumen

df_j = jumlah dokumen yang mengandung istilah t_j

Terdapat beberapa macam perhitungan *Similarity Coeficient* yaitu menggunakan rumus *cosine measure* dan *normalized cosine measure*.

$$nw_{ij} = \frac{freq_{ij}}{\max_k freq_{ik}} * nidf_j \qquad (2)$$

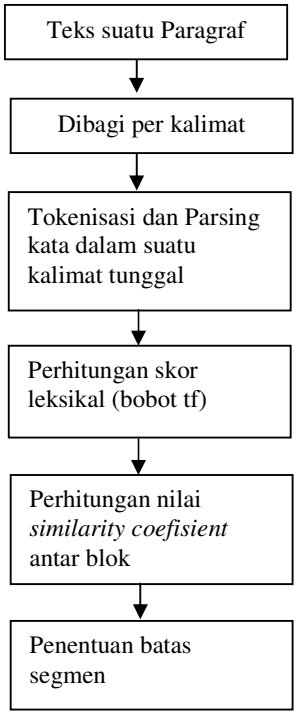
$$nd_{ij} = ntf_{ij} * nidf_j \qquad (3)$$

$$ntf_{ij} = \frac{tf_{ij}}{\max_k tf_{ik}} \qquad (4)$$

$$nidf_j = \frac{\log(d) - \log(df_j)}{\log(d)} = 1 - \frac{\log(df_j)}{\log(d)} \qquad (5)$$

4. Metodologi Perancangan

Model yang akan dikembangkan dalam penentuan topik dalam suatu paragraf dalam penelitian dapat dilihat dalam gambar 1 di bawah ini.



Gambar 1. Langkah-langkah penentuan topik suatu paragraf

Kesulitan dalam penentuan kalimat adalah tanda titik yang ada. Dalam penelitian ini kajian permasalahan dibatasi pada tidak diacuhkannya format penanda titik. Titik terkadang tidak menentukan akhir suatu kalimat. Titik dapat digunakan untuk identifikasi akhir jalan, gelar dan lainnya. Gambar 1 adalah alur proses dari pencarian topik secara otomatis dalam suatu paragraf. Langkah pertama adalah dilakukannya penelusuran (*scanning*) terhadap paragraf untuk memilah paragraf tersebut perkalimatnya.

 ISBN: 979-756-061-6

Contoh:
"Struktur MIS dipengaruhi aktifitas manajemen dan fungsi organisasi. Kebutuhan informasi bervariasi tergantung tingkat penunjang aktifasi management: rencana strategis, kontrol manager, atau kontrol operasi. Mereka juga bervariasi tergantung tingkatan struktur penunjang keputusan. Setiap fungsi organisasi ditunjang oleh sistem informasi mempunyai kebutuhan proses informasi unik seperti pada umumnya."

Sehingga pembagian per kalimat menjadi:
Kalimat pertama:
"Struktur MIS dipengaruhi aktifitas manajemen dan fungsi organisasi"
Kalimat kedua:
"Kebutuhan informasi bervariasi tergantung tingkat penunjang aktifasi management: rencana strategis, kontrol manager, atau kontrol operasi"
Kalimat ketiga:
"Mereka juga bervariasi tergantung tingkatan struktur penunjang keputusan"
Kalimat keempat:
"Setiap fungsi organisasi ditunjang oleh sistem informasi mempunyai kebutuhan proses informasi unik seperti pada umumnya"

Setelah dilakukan pembagian blok per kalimat, kemudian dilakukan proses *parsing* dan pembuatan indeks masing-masing kata beserta letak kalimat keberapa pada samping bobot. Bobot dihitung berdasarkan besar frekuensi yang muncul dalam setiap kalimat dan seberapa seringnya kata tersebut muncul secara keseluruhan pada satu paragraf tersebut. Semakin banyak kata tersebut muncul pada setiap kalimat, maka nilai bobot kata tersebut akan besar, tetapi bila kata tersebut muncul pada beberapa kalimat maka bobot kata tersebut akan berkurang (*inverse document frequency*). Sebelum di indeks maka ada kata-kata yang akan dibuang dari kalimat tersebut (*stoplist*), pembuangan dilakukan karena sering munculnya kata tersebut sehingga diasumsikan sebagai kata yang umum dan tidak berguna (seperti atau, dan, dia, saya, oleh dan lain-lain).

Proses selanjutnya adalah perhitungan nilai *Similarity Coeficient* (SC) dengan menggunakan metoda *normalized cosine measure* (pers 3, 4, dan 5). Pada model ini dilakukan normalisasi terlebih dahulu terhadap nilai *term frequency* (tf) maupun terhadap nilai *inverse document frequensi* nya (idf). Banyak terdapat model untuk mencari nilai kemiripan tersebut (selain model pada rumus 1), model lain diantaranya adalah:

Model **dot product**

$$SC(Q,Di) = (Di)(Q) \qquad (6)$$

Model **Dice**

$$SC(Q,Di) = 2(Di)(Q)/(Di2)(Q2) \qquad (7)$$

Model **Jaccard**

$$SC(Q,Di) = (DiQ)/((Di)2 + Q2 - |Di||Q|) \qquad (8)$$

Selanjutnya setiap kata setiap kalimat tersebut dimasukkan ke dalam blok token tiap posisinya. Beserta bobot *term frequency* dan *inverse document frequency* nya. Contoh pembentukan blok dapat dilihat pada tabel 1 di bawah ini.

Tabel 1. Contoh TfIdf *term* dalam blok

Token	***TfIdf***		
	Blok 1	***Blok 2***	***Blok 3***
A	2	1	0
B	1	1	1
C	2	1	0
D	1	1	0
E	1	2	1
F	0	0	2
G	0	0	1

Dalam tabel tersebut menyatakan bahwa kata/token A terdapat pada kalimat ke 1 dan ke 2 dengan frekuensi TfIdf 2 dan 1, begitu juga seterusnya. Kemudian perhitungan kemiripan antar kalimat pada Blok 1, Blok 2 dan Blok 3 adalah sebagai berikut:

$$\cos(B1, B2) = \frac{(2)(1)+(1)(1)+(2)(1)+(1)(1)+(1)(2)}{\sqrt{(2^2+1^2+2^2+1^2+1^2)(1^2+1^2+1^2+1^2+2^2)}}$$

$$= \frac{8}{\sqrt{(11)(8)}} = 0.8528$$

Dengan cara yang sama diperoleh nilai kesamaan antara b2 (Blok 2) dan b3 (blok 3). Selanjutnya dilakukan proses identifikasi batas segmentasi diantara nilai-nilai SC tiap-tiap relasi kalimat tiap paragraf tersebut.

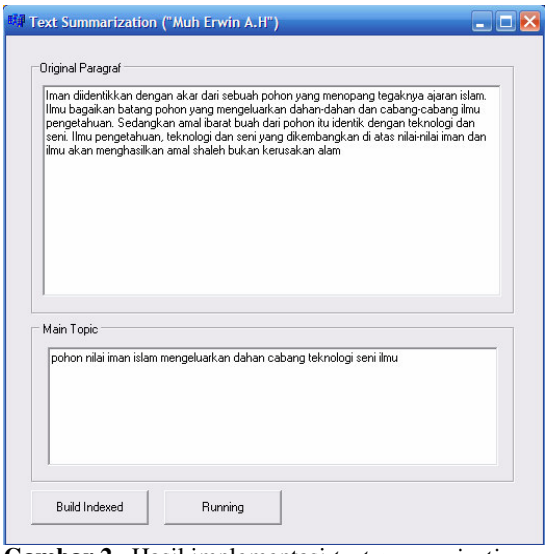
Dalam penelitian sebelumnya Tannen (1989), Halliday & Hasan (1976), Walker (1991), dan Hearst [MAN04] menemukan bahwa perulangan suatu term atau kata dalam suatu kalimat dalam suatu paragraf menunjukkan indikasi sangat penting dalam menentukan struktur topik paragraf. Blok-blok dibangun berdasar kalimat-kalimat yang ada di dalam paragraf yang akan dicari topik nya. Identifikasi keterkaitan secara leksikal adalah dengan membandingkan pasangan blok-blok teks berurutan, kemudian dihitung besar kemiripannya antara dua blok teks. Semakin besar kemiripannya maka topik cerita antar kalimat saling berkelanjutan. [MAN04]

Penentuan nilai segmen blok dihitung dengan menggunakan nilai *similarity coefisient* antar dua kalimat di atas. Hasil nilai segmen kemudian akan di ranking secara descending dan kemudian ditentukan segmen n terbesar dari nilai paling atas.

ISBN: 979-756-061-6

5. Implementasi

Penelitian dilakukan terhadap sejumlah teks paragraf dari suatu surat kabar lokal Kedaulatan Rakyat. Tema yang diambil dari kasus uji tersebut adalah pendidikan dan budaya. 20 Sampel cerita diambil, dimana setiap sampel akan dipilah-pilah menurut paragraf-paragrafnya. Setelah itu tiap paragraf diberikan *judgment* terhadap intisarinya secara manual terlebih dahulu oleh penulis dan rekan penulis. Implementasi dilakukan dengan mencocokkan hasil topik automatis oleh sistem dan hasil topik yang telah ditentukan oleh penulis untuk masing-masing paragrafnya.



Gambar 2. Hasil implementasi *text summarization* terhadap sebuah contoh paragraf masukan

6. Penutup dan Kesimpulan

Perangkat lunak yang dibangun menggunakan kakas pemrograman Borland Builder C++ versi 6 baru merupakan intisari kasus permasalahan saja. Perangkat penentuan topik secara otomatis dalam suatu paragraf akan dikembangkan lagi untuk dapat memprediksi topik dari seluruh pustaka yang ada. Hasil yang dicapai sudah cukup menggambarkan kesimpulan topik yang diinginkan walaupun pada beberapa bagian masih terlihat kejanggalan susunan pembentukan kata tersebut. Kesimpulan yang didapat untuk penelitian selanjutnya adalah mengembangkannya dengan melihat susunan pembentuk kata tata bahasa (*grammar*) sehingga intisari topik memang menggambarkan susunan kata yang alami bukan berdasar penggabungan yang tidak sesuai dengan tata bahasa yang telah ditentukan oleh suatu bahasa tertentu. Selain itu penentuan kemiripan kata dalam suatu kalimat dalam suatu paragraf dapat ditambahkan thesaurus (kamus) kata yang memiliki keterkaitan kata yang berelasi (*related term*), kata berdekatan arti (*narrower term*), dan kata yang berjauhan arti (*broader term*), sehingga kata-kata yang jelas memiliki arti dan semantik yang berjauhan akan langsung dieliminasi dari pencarian kesesuaian kata tersebut. Hal ini tentu akan lebih memperingkas waktu dan hasil yang dicapai akan lebih baik.

Hasil yang didapat menunjukkan dari 20 sampel cerita yang diambil menunjukkan 60 prosen berhasil mengenali dan mengidentifikasi topik suatu paragraf walaupun susunan kata pembentuk topik tersebut terkadang tidak sesuai dengan kaidah tata bahasa yang dimaksud. 40 prosen tidak mengenali topik suatu paragraf. Penentuan nilai 40 prosen tersebut didasarkan pada pembentukan kata-kata yang sangat tidak sesuai dengan apa yang dimaksud (sesuai dengan jugdment pengguna sebelumnya). Sehingga dari 20 kali percobaan dijumpai 8 kali kegagalan. Maka dapat disimpulan bahwa banyaknya kegagalan ini mungkin dikarenakan oleh penentuan rumusan nilai batas blok antar nilai SC yang didapat antar kalimat tersebut.

Daftar Pustaka

[FRA92] Frakes William and Ricardo Baeza-Yates; *Information Retrieval Data Structure and Algorithms*, Prentice Hall, 1992.

[GRO98] Grossman David, and Ophir Frieder, *Information Retrieval : Algorithms and Heuristics*, Kluwer Academic Publisher, 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/publication.shtml

[MIY90] Miyamoto, Sadaki; *Fuzzy sets in Information Retrieval and cluster analysis*; Kluwer Academic Publisher; London, 1990.

[ERW04] Erwin, Muhammad, "*Relevance feedback pada sistem temu kembali informasi menggunakan algoritma genetika*", Thesis Magister Informatika ITB, 2004.

[MAN99] Mandala Rila, Takenobu Takunaga, Hozumi Tanaka. "*Query expansion using heterogenous thesauri*. "Proceeding of Information Processing and Management. 1999.

[MAN00] Mandala Rila, Takenobu Takunaga, Hozumi Tanaka. "*The exploration and Analysis of Using Multiple Thesaurus types for Query Expansion in Information Retrieval*". Journal of Information Processing. 2000.

[MAN02] Mandala Rila, "*Sistem Temu-kembali informasi dengan menggunakan model probabilistik*" Jurnal Informatika, ITB, Bandung, 2002.

[MAN04] Mandala, Rila. Andreas Prasetia, Rinaldi Munir, Harlili. Sistem pengidentifikasi otomatis keterkaitan topik antar paragraf dalam dokumen ekspositori. Prosiding SNATI 2004.

[SET03] Setiawan, Kuswara. "*Paradigma Sistem Cerdas*". BayuMedia Publishing, Malang. Jawa Timur, 2003.