HIERARCHICAL CLUSTERING UNTUK APLIKASI *AUTOMATED TEXT INTEGRATION*

Gregorius S. Budhi[1] ; Arlinah I. Rahardjo[2] ; Hendrawan Taufik[3]
Universitas Kristen Petra Jurusan Teknik Informatika
Jalan Siwalankerto 121-131 Surabaya 60236, Jawa Timur, Indonesia
E-mail: greg@petra.ac.id , arlinah@petra.ac.id

ABSTRAK
Membaca beberapa dokumen yang membahas topik yang sama memerlukan waktu yang lama. Peneliti mencoba membuat aplikasi Automated Text Integration yang dapat menghasilkan integrasi dari beberapa dokumen yang berbeda dengan topik bahasan yang sama. Aplikasi ini memberi kemudahan kepada pembaca dalam menggali informasi pada dokumen – dokumen tersebut. Teknik Data Mining Hierarchical Clustering digunakan untuk mengintegrasikan dokumen – dokumen yang berbeda itu. Metode perhitungan bobot kalimat, yang dimodifikasi dari penelitian sebelumnya [5], berfungsi untuk menghitung nilai bobot dari setiap kalimat. Nilai bobot ini digunakan sebagai dasar penggabungan cluster. Terakhir Cosine Distance digunakan untuk menghitung similarity (tingkat kesamaan) antar dokumen – dokumen yang akan diintegrasi. Dari hasil survei terhadap 100 orang responden, sebanyak 78% responden mengatakan bahwa integrasi dokumen yang dihasilkan telah benar. Selain itu, hasil integrasi yang baik akan didapat bila jenis dokumen yang diintegrasi bertipe eksposisi.
Keywords: *Integrasi Dokumen, Hierarchical Clustering, Cosine Distance*

1. PENDAHULUAN

Salah satu cara untuk memperoleh informasi seimbang adalah dengan membaca beberapa dokumen yang membahas topik yang sama. Namun hal ini menyulitkan pembaca untuk menangkap topik bahasan utama dari dokumen - dokumen tersebut karena harus mengingat – ingat isi dokumen yang telah dibaca sebelumnya. Pembaca harus mengintegrasikan dahulu dokumen – dokumen yang dia baca didalam pikirannya sebelum dapat merangkum maksud dan topik utama dokumen – dokumen tersebut secara keseluruhan.

Pada penelitian ini peneliti mencoba membuat aplikasi *Automated Text Integration* yang dapat menghasilkan integrasi dari beberapa dokumen elektronik yang berbeda dengan topik bahasan yang sama secara otomatis. Proses integrasi akan menghasilkan dokumen baru yang mengandung semua bagian dari dokumen – dokumen awal, namun memiliki susunan antar kalimat serta antar paragraf yang berbeda. Perbedaan ini karena saat proses integrasi topik – topik bahasan yang serupa (*similar*) dari semua dokumen dikumpulkan menjadi satu paragraf dan disusun ulang kalimat per kalimat sesuai dengan besarnya kesamaan (*similiarity*) antar kalimatnya. Dengan membaca hasil integrasi diharapkan pembaca dapat terbantu dalam menyerap informasi penting yang ada dalam kumpulan dokumen yang berbeda dan tidak perlu lagi membaca sekumpulan dokumen satu per satu.

2. TINJAUAN PUSTAKA

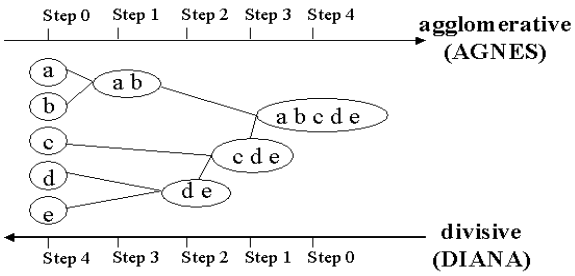
2.1. *Hierarchical Clustering*

Pada algoritma *clustering*, *data* akan dikelompokkan menjadi *cluster-cluster* berdasarkan kemiripan satu *data* dengan yang lain. Prinsip dari *clustering* adalah memaksimalkan kesamaan antar anggota satu *cluster* dan meminimumkan kesamaan antar anggota *cluster* yang berbeda [6, 9].

Kategori algoritma *clustering* yang banyak dikenal adalah *Hierarchical Clustering*. *Hierarchical Clustering* adalah salah satu algoritma *clustering* yang dapat digunakan untuk meng-*cluster* dokumen *(document clustering*). Dari teknik *hierarchical clustering*, dapat dihasilkan suatu kumpulan partisi yang berurutan, dimana dalam kumpulan tersebut terdapat:

a. *Cluster – cluster* yang mempunyai poin – poin individu. *Cluster – cluster* ini berada di level yang paling bawah.
b. Sebuah *cluster* yang didalamnya terdapat poin – poin yang dipunyai semua *cluster* didalamnya. *Single cluster* ini berada di *level* yang paling atas.

Hasil keseluruhan dari algoritma *hierarchical clustering* secara grafik dapat digambarkan sebagai *tree*, yang disebut dengan *dendogram*. *Tree* ini secara grafik menggambarkan proses penggabungan dari *cluster – cluster* yang ada, sehingga menghasilkan *cluster* dengan level yang lebih tinggi [9]. Gambar 1 adalah contoh *dendogram*.



Gambar 1. *Dendogram* [6]

2.1.1. Agglomerative Hierarchical Clustering

Metode ini menggunakan strategi disain *Bottom-Up* yang dimulai dengan meletakkan setiap obyek sebagai sebuah *cluster* tersendiri (*atomic cluster*) dan selanjutnya menggabungkan *atomic cluster – atomic cluster* tersebut menjadi *cluster* yang lebih besar dan lebih besar lagi sampai akhirnya semua obyek menyatu dalam sebuah *cluster* atau proses dapat pula berhenti jika telah mencapai batasan kondisi tertentu [6]. Metode *Agglomerative Hierarchical Clustering* yang digunakan pada penelitian ini adalah metode *AGglomerative NESting* (AGNES). Cara kerja AGNES dapat dilihat pada gambar 1.

Adapun ukuran jarak yang digunakan untuk menggabungkan dua buah obyek cluster adalah *Minimum Distance* [6], yang dapat dilihat pada persamaan 1.

$$d_{\min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} |p - p'| \quad ...(1)$$

Dimana |p – p'| jarak dua buah obyek p dan p'.

2.2. Algoritma *Cosine Distance*

Metode cosine distance merupakan metode yang digunakan untuk menghitung *similarity* (tingkat kesamaan) antar dua buah obyek [6]. Pada penelitian ini obyek Berikut adalah persamaan dari metode *Cosine Distance* :

$$Similiarity(v_1, v_2) = \frac{v_1 \bullet v_2}{|v_1||v_2|} \quad \text{..... (2)}$$

Pada penelitian ini obyek v1 dan v2 adalah dua buah dokumen yang berbeda.

2.3. Proses Parsing, Stemming dan Stopword Removal

Dalam bidang tata bahasa dan linguistik, *parsing* adalah sebuah proses untuk menjadikan sebuah kalimat menjadi lebih bermakna atau berarti dengan cara memecah kalimat tersebut menjadi kata-kata atau frase – frase [4].

Stemming adalah proses pemetaan dan penguraian berbagai bentuk (*variants*) dari suatu kata menjadi bentuk kata dasarnya. Proses *stemming* digunakan di dalam proses *Information Retrieval* (pencarian informasi) untuk meningkatkan kualitas informasi yang didapatkan [4].

Stopwords removal adalah sebuah proses untuk menghilangkan kata yang 'tidak relevan' pada hasil *parsing* sebuah dokumen teks dengan cara membandingkannya dengan *Stoplist* (*Stopword list*) yang ada [4]. Contoh dari *Stopword* misalnya, kata sambung, artikel dan preposisi.

2.4. Bobot Relasi antar kalimat

Bobot relasi antara dua kalimat adalah sama dengan jarak antara kedua kalimat tersebut. Konsekuensinya adalah bila bobot relasi antara dua kalimat tertentu lebih kecil dari yang lain, maka jarak keduanya juga lebih dekat [5, 8]. Secara formal, misal terdapat n kalimat P={S_1, S_2, …, S_n}, maka bobot relasi antara dua kalimat S_i dan S_j dapat dilihat pada persamaan 3.

$$R(S_i, S_j) = \begin{cases} \dfrac{(j-i)^2}{\alpha(S_i, S_j) \times W(S_j)}, & i < j \\ \infty, & i \geq j \end{cases} \quad \text{.....(3)}$$

dimana i, j adalah letak kalimat ke i dan j; $\alpha(S_i, S_j)$ adalah jumlah kata yang sama atara S_i dan S_j setelah *stopword* yang ada dihilangkan ; dan $W(S_j)$ adalah bobot kalimat ke j.

Pada penelitian sebelumnya [5] letak kalimat ke i dan j diukur hanya pada satu paragraf saja. Pada penelitian ini definisi tersebut diubah, yaitu: i dan j adalah nomor urut kalimat pada gabungan dokumen yang disusun secara berurutan berdasarkan relasi antar dokumen, yang diukur menggunakan *Cosine Distance* (persamaan 2).

2.5. Bobot Kalimat

Bobot Kalimat adalah sebuah nilai senuah kalimat yang mengindikasikan sebeberapa penting arti kalimat tersebut pad sebuah paragraf. Semakin tinggi nilai kalimatnya semakin penting pula artinya dalam paragraf. Proses *Parsing, Stemming* dan *Stopword Removal* harus dikerjakan terlebih dahulu sebelum proses perhitungan bobot kalimat ini dilakukan [4].

Perhitungan bobot kalimat ini berbasis pada [8] dan telah dimodifikasi pada penelitian sebelumnya [5]. Ada empat macam bobot kalimat yang digunakan pada penelitan sebelumnya yaitu:

- W1 → Banyaknya kata yang sama antara kalimat yang dihitung dengan daftar kata kunci *(keyword)* pada dokumen tempat kalimat tersebut berada.
- W2 → Nilai yang ditentukan dari kemunculan kata – kata didalam kalimat terhadap pemakaian kata – kata tersebut pada dokumen tempat kalimat berada.
- W3 → Nilai ini ditentukan oleh posisi dimana kalimat tersebut berada terhadap paragrafnya. Berdasarkan kaidah Deduktif – Induktif bahasa Indonesia ada 2 macam

nilai yang dipakai disini, yaitu: Bila kalimat tersebut berada pada awal / akhir paragraf memiliki bobot 2, sementara bila tidak memiliki bobot 1.

W4 → Menghitung banyaknya relasi sebuah kalimat dengan kalimat – kalimat lain pada dokumen yang sama.

Bobot Kalimat total dapat dilihat pada persamaan 4.

$$W(S_j) = W1(S_j) + W2(S_j) + W3(S_j) + W4(S_j) \quad \ldots (4)$$

dimana j adalah kalimat ke-j dari total n kalimat.

Untuk penelitian kali ini perhitungan bobot kalimat ini dimodifikasi kembali agar sesuai kebutuhan pada penelitian ini. Pemikiran dari modifikasi ini adalah:

- Pada penelitian terdahulu [5] proses hanya diterakan pada satu dokumen saja, oleh sebab itu bobot dari kalimat cukup dihitung terhadap sebuah dokumen saja.
- Pada penelitian ini ada beberapa dokumen yang digabungkan, untuk itu perlu diperhitungkan bahwa bobot sebuah kalimat tidak hanya diukur terhadap kalimat lain pada dokumen yang sama melainkan juga terhadap kalimat lain di dokumen yang berbeda yang akan diintegrasikan.

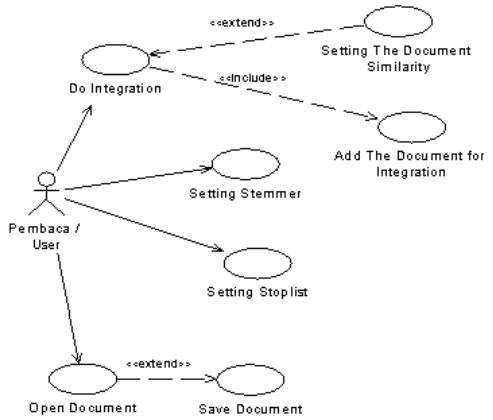
Dari pemikiran diatas, peneliti akhirnya memutuskan bahwa bobot kalimat pada persamaan 4 perlu dimodifikasi dengan sebuah bobot kelima. Bobot kelima ini (W5) merepresentasikan seberapa penting sebuah kalimat dibandingkan dengan kalimat – kalimat lain yang terdapat pada semua dokumen yang akan diintegrasikan. Persamaan hasil modifikasi dapat dilihat pada persamaan 5.

$$W(S_j) = W1(S_j) + W2(S_j) + W3(S_j) + W4(S_j) + W5(S_j) \quad \ldots\ldots\ldots\ldots (5)$$

dimana W5 adalah Banyaknya kata kunci (keyword) yang sama antara kalimat yang dihitung dengan daftar kata kunci pada semua dokumen yang akan diintegrasikan. Asumsinya adalah semakin banyak kata pada kalimat tersebut sama dengan daftar kata kunci, semakin penting keberadaan kalimat tersebut pada dokumen hasil integrasi.

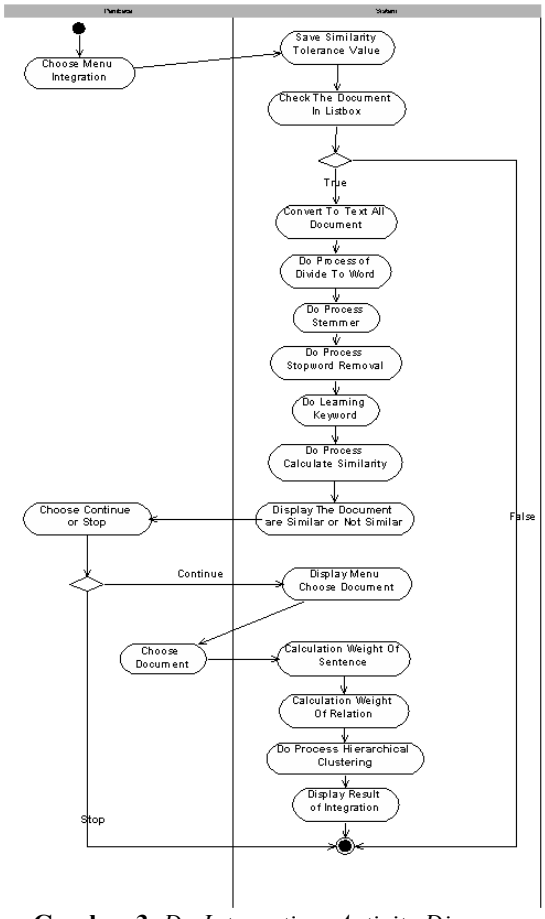
3. DESAIN APLIKASI *AUTOMATED TEXT INTEGRATION*

Desain aplikasi *Automated Text Integration* dapat dilihat pada diagram *Use Case* pada gambar 2



Gambar 2. Diagram *Use Case* Aplikasi

Inti dari aplikasi ini adalah *Do Integration*, dimana pada *use case* ini proses integrasi beberapa dokumen yang dipilih dilakukan. Diagram *activity* dari *use case* ini dapat dilihat pada Gambar 3.



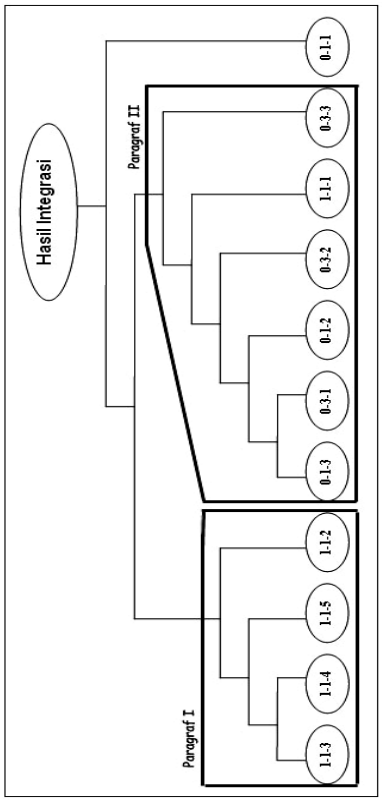
Gambar 3. *Do Integration Activity Diagram*

Penjelasan dari diagram *activity* pada gambar 3 adalah sebagai berikut:

- Sebelum proses ini, *user* harus menentukan terlebih dahulu dokumen – dokumen mana yang akan diintegrasikan pada menu 'Open Dokumen'. Proses ini secara otomatis akan menulis informasi nama dan *path* dokumen –

dokumen tersebut pada *listbox* dokumen. Format dokumen yang dapat dipilih adalah *.doc dan *.txt.

- Setelah *user* memilih menu *'Integration'*, aplikasi akan meminta user mengisikan nilai batas terendah *similarity* antar dokumen yang diijinkan oleh *user* untuk dokumen – dokumen yang akan diintegrasikan.
- Selanjutnya bila *listbox* dokumen terisi, aplikasi akan merubah semua dokumen yang ada kedalam bentuk teks, merubahnya menjadi sekumpulan kalimat dan kata – kata yang berurutan (*divide to word / parsing*), melakukan proses *stemming*, *stopword removal*, menandai kata – kata mana saja yang merupakan *keyword*, dan menghitung *similarity* antar dokumen dengan persamaan 2.
- Selanjutnya aplikasi akan menunjukkan *list similarity* antar dokumen dan memberi tanda bila *similarity* tersebut dibawah nilai yang telah ditentukan. Bila *user* memilih melanjutkan proses dengan memilih *'continue'*, aplikasi akan menyusun dokumen – dokumen tersebut secara berurutan sesuai dengan *level similarity*-nya.
- Langkah berikutnya aplikasi akan menghitung bobot kalimat (*Weight Of Sentence*) dan bobot relasi antar kalimat (*Weight Of Relation*). Bobot relasi antar kalimat ini yang akan dipakai untuk mengintegrasikan dokumen menggunakan metode *AGglomerative NESting (AGNES)*.
- Pada proses integrasi, awalnya semua kalimat pada semua dokumen dianggap sebagai *atomic cluster – atomic cluster*. Selanjutnya secara bertahap *cluster – cluster* tersebut akan disatukan menggunakan aturan *Minimum Distance* pada persamaan 1. Setelah semua kalimat telah tergabung menjadi sebuah *cluster*, dilakukan proses untuk memecah *cluster* tersebut menjadi paragraf – paragraf. Caranya adalah, kalimat – kalimat yang bergabung terlebih dahulu menjadi *cluster – cluster* besar dianggap sebagai sebuah paragraf tersendiri. Asumsinya, bila secara *natural* kalimat – kalimat tersebut bergabung, dapat dianggap kalimat – kalimat tersebut memiliki *similarity* yang cukup tinggi dan membahas topik bahasan yang sama. Agar lebih jelas, proses integrasi ini dapat dilihat pada gambar 4. Sementara untuk memproses kalimat – kalimat tersisa yang tidak mau bergabung kedalam *cluster – cluster* besar, dipakai aturan sebagai berikut:
  - Bila hanya 1 kalimat (seperti kalimat no. 0-1-1 pada gambar 4) akan digabungkan pada paragraf terakhir.
  - Bila lebih dari satu kalimat, kalimat – kalimat yang tersisa tersebut akan dipaksakan bergabung menjadi satu paragraf tersendiri.



Gambar 4. Proses Integrasi menggunakan AGNES

- Langkah terakhir adalah menyuguhkan hasil integrasi kepada *user* dalam bentuk tampilan teks. *User* kemudian dapat memilih untuk menyimpan hasil integrasi kedalam file *.doc atau *.txt.

4. PENGUJIAN APLIKASI

4.1. Pengujian Hasil Integrasi

Untuk membandingkan hasil integrasi dengan dokumen aslinya. Dua buah dokumen pendek aslinya pada gambar 5 dan 6 digabungkan dan pada gambar 7 dapat dilihat hasil integrasinya. Pada gambar 4 dapat dilihat bagaimana proses penggabungannya. Gambar 5 menjadi dokumen ke - 0 dan gambar 6 adalah dokumen ke – 1.

"Samsung Pecahkan Rekor Layar *LCD* Tertipis"

Inovasi tiada henti terus dipertunjukkan Samsung Electronics. Kali ini perusahaan elektronik asal Korea Selatan tersebut memperkenalkan layar LCD yang diklaim tertipis di dunia. Ukuran ketebalan LCD 0,74 mm mematahkan rekor layar LCD tertipis yang sebelumnya juga dibuat Samsung, setebal 0,82 mm.

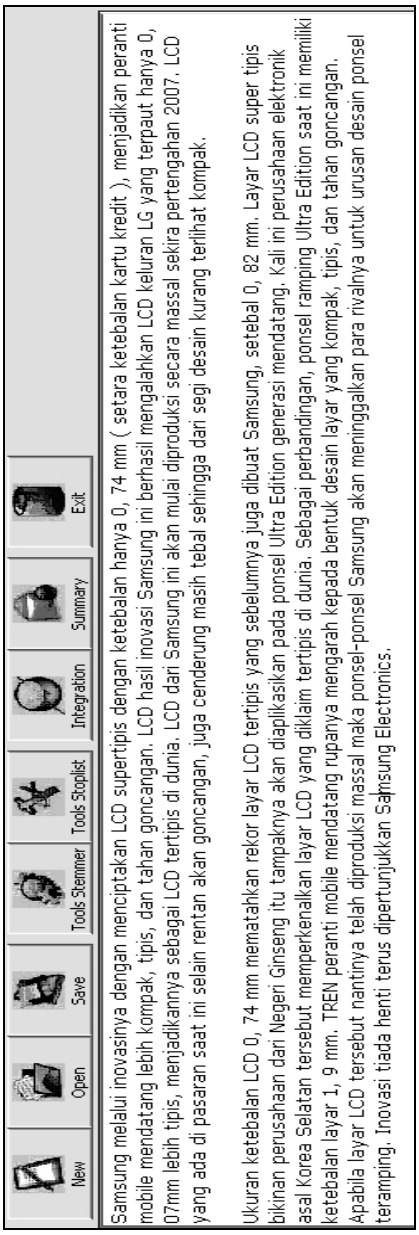
Layar LCD super tipis bikinan perusahaan dari Negeri Ginseng itu tampaknya akan diaplikasikan pada ponsel Ultra Edition generasi mendatang. Sebagai perbandingan, ponsel ramping Ultra Edition saat ini memiliki ketebalan layar 1,9 mm. Apabila layar LCD tersebut nantinya telah diproduksi massal maka ponsel-ponsel Samsung akan meninggalkan para rivalnya untuk urusan desain ponsel teramping.

Gambar 5. Dokumen asal ke - 0

"*LCD* Tertipis di Dunia"

TREN peranti *mobile* mendatang rupanya mengarah kepada bentuk desain layar yang kompak, tipis, dan tahan goncangan. LCD yang ada di pasaran saat ini selain rentan akan goncangan, juga cenderung masih tebal sehingga dari segi desain kurang terlihat kompak. Samsung melalui inovasinya dengan menciptakan LCD supertipis dengan ketebalan hanya 0,74 mm (setara ketebalan kartu kredit), menjadikan peranti *mobile* mendatang lebih kompak, tipis, dan tahan goncangan. LCD hasil inovasi Samsung ini berhasil mengalahkan LCD keluran LG yang terpaut hanya 0,07mm lebih tipis, menjadikannya sebagai LCD tertipis di dunia. LCD dari Samsung ini akan mulai diproduksi secara massal sekira pertengahan 2007.

Gambar 6. Dokumen asal ke - 1



Gambar 7. Hasil Integrasi

4.2. Pengujian Dalam Bentuk *Survey*

Pengujian ini dilakukan dengan cara meminta bantuan 100 orang responden umum untuk membaca dokumen – dokumen asal dan dokumen hasil integrasi, kemudian menjawab 3 pertanyaan berikut:

1. Menurut anda, apakah kata-kata pada dokumen hasil integrasi tersebut telah terorganisir dengan baik (tiap paragraf memberikan arti yang jelas dan dapat dipahami) ? A. Ya B. Tidak
2. Menurut anda, apakah dokumen hasil integrasi tersebut telah memberikan gambaran secara umum dari keseluruhan dokumen yang ada sebelumnya ? A. Ya B. Tidak
3. Menurut anda, apakah dokumen hasil integrasi dapat memberikan informasi - informasi penting yang terdapat pada dokumen sebelumnya secara jelas? A. Ya B. Tidak

Hasil survey dapat dilihat pada tabel 1.

Tabel 1. Hasil *Survey*

Integrasi 3 Dokumen tentang "*Data Mining*"		
	A (%)	**B (%)**
Pertanyaan 1	77.00	23.00
Pertanyaan 2	83.00	17.00
Pertanyaan 3	85.00	15.00
Jenis Kelamin Responden	**Laki-laki (%)**	**Perempuan (%)**
	73	27
Integrasi 2 dokumen tentang *"High-Speed Downlink Packet Access (HSDPA)"*		
	A (%)	**B (%)**
Pertanyaan 1	78.00	22.00
Pertanyaan 2	72.00	28.00
Pertanyaan 3	78.00	22.00
Jenis Kelamin Responden	**Laki-laki (%)**	**Perempuan (%)**
	64	36

Kedua jenis dokumen yang dipakai menjadi bahan *survey* bertipe eksposisi, yaitu dokumen yang berusaha menjelaskan suatu prosedur atau proses, memberikan definisi, menerangkan, menjelaskan, menafsirkan gagasan, menerangkan bagan atau tabel, atau mengulas sesuatu kepada pembaca.

Sementara untuk dokumen berbentuk naratif seperti cerita rakyat, tidak disertakan dalam *survey*, karena peneliti sendiri telah melihat adanya kerancuan pada jalan cerita pada dokumen hasil integrasinya. Hal ini selalu terjadi pada beberapa uji coba pada beberapa topik dokumen naratif, seperti "Timun Emas", "Sangkuriang", "Jack dan Kacang Polong" dan lain – lainnya. Oleh karena itu dapat disimpulkan bahwa proses integrasi ini tidak cocok untuk dokumen yang berjenis naratif.

4.3. Pengujian Kecepatan Proses

Pengujian kecepatan proses aplikasi *Automated Text Integration* ini dilakukan pada spesfikasi *hardware* dan *software* berikut ini, *Processor: Pentium* IV 1600 *MHz*; *Memory:* 512

Mbyte; *HardDisk:* 40 *Gigabyte* dan *Operating System: Windows XP Professional.* Hasil pengujian dapat dilihat pada tabel 2.

Tabel 2. Hasil Pengujian Kecepatan Proses

No. Integrasi	No. Dokumen	Judul Dokumen	Ukuran Dokumen		Ukuran Hasil Integrasi		Total Waktu Proses	Waktu Hierarchical Clustering
			Jumlah Paragraf	Jumlah Kalimat	Jumlah Paragraf	Jumlah Kalimat		
1	1	Mewaspadai Penyakit Lingkungan Akibat Radiasi Elektromagnetik	6	15	4	47	08m 03s 000ms	01m 08s 844ms
	2	Pengaruh Gelombang Elektromagnetik Ponsel pada Kesehatan	5	10				
	3	Radiasi di Sekitar Kita	6	22				
2	4	algoritma genetik 1	1	5	2	13	01m 56s 406ms	00m 03s 906ms
	5	algoritma genetik 2	2	8				
3	6	HTML 1	2	11	4	37	04m 07s 312ms	00m 35s 656ms
	7	HTML 2	2	10				
	8	HTML 3	2	16				

5. KESIMPULAN

- Berdasarkan hasil uji coba aplikasi dan survei yang dilakukan kepada 100 orang responden didapat bahwa hasil integrasi yang diperoleh dari proses tersebut memberikan hasil yang baik. Baik disini berarti dapat merepresentasikan secara jelas isi dari dokumen asal dan juga dapat memberikan informasi – informasi penting yang terdapat pada dokumen asal.
- Hasil dari proses integrasi dapat disimpan dalam bentuk *file*. Hal ini dapat membantu user / pembaca untuk membaca ulang dokumen hasil integrasi tersebut tanpa harus menjalankan kembali proses integrasi.
- Berdasarkan hasil survei yang dilakukan kepada 100 orang responden disimpulkan bahwa aplikasi ini cocok untuk dokumen yang bertipe eksposisi.
- Sementara, dari hasil pengujian dan pengamatan oleh peneliti dapat disimpulkan bahwa aplikasi tidak cocok untuk dokumen bertipe naratif. Hal ini disebabkan karena pada saat dokumen yang bertipe naratif diintegrasi, maka akan selalu terjadi kerancuan terhadap jalan cerita yang dimiliki oleh dokumen hasil integrasi.
- Hasil pengujian waktu proses menunjukkan bahwa semakin banyak dokumen, paragraf, dan kalimat yang diproses maka akan membutuhkan waktu proses yang semakin besar pula. Namun waktu proses tersebut, yang hanya dalam kisaran menit, masih dapat dianggap wajar. Dari sini dapat disimpulkan bahwa aplikasi ini dapat digunakan pada dunia nyata.

DAFTAR PUSTAKA

[1] Akhadiah, Sabarti, Maidar M. K. Arsjad dan Sakura Ridwan, *Buku Materi Pokok : Bahasa Indonesia*, Jakarta: Penerbit Karunika Jakarta UT. 1986.

[2] Arifin, E. Zaenal, dan Amran Tasai, *Cermat Berbahasa Indonesia Untuk Perguruan Tinggi*, Jakarta: Penerbit Akademika Pressindo, 2000.

[3] Garcia, E., **"**An information retrieval tutorial on cosine similarity measures, dot products and term weight calculations", 2006, http://www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html (January, 2007)

[4] Gregorius S. Budhi, Ibnu Gunawan dan Ferry Yuwono, "Algortima Porter Stemmer For Bahasa Indonesia Untuk Pre-Processing Text Mining Berbasis Metode Market Basket Analysis", PAKAR Jurnal Teknologi Informasi Dan Bisnis vol. 7 no. 3 November, 2006.

[5] Gregorius S. Budhi; Rolly Intan, Silvia R. dan Stevanus R. R., "Indonesia Automated Text Summarization". Proceeding ICSIIT 2007. , 26 - 27 July 2007.

[6] Han, Jiawei and Micheline Kamber, *Data Mining: Concepts and Techniques,* Morgan Kaufmann, 2001.

[7] Pusat Pembinaan & Pengembangan Bahasa Departemen Pendidikan dan Kebudayaan Republik Indonesia, *Pedoman umum ejaan bahasa Indonesia yang disempurnakan*. Jakarta: Balai Pustaka, 1999.

[8] Sjobergh, Jonas, and Kenji Araki, *Extraction based summarization using a shortest path algorithm*. Sweden: KTH Nada, 2005.

[9] Steinbach, M., G. Karypis and Vipin Kumar, *A comparisont of document clustering techniques*, Minnesota: University of Minnesota, Department of Computer Science and Engineering, 2000, http://glaros.dtc.umn.edu/gkhome/fetch/papers/doccluster.pdf (January, 2007)