Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 3, No. 1, Januari 2019, hlm. 325-331 http://j-ptiik.ub.ac.id

Peringkasan Multi-Dokumen Berbasis *Clustering* pada Sistem Temu Kembali Berita *Online* Menggunakan Metode *K-Means*

Amalia Kusuma Akaresti[1], Mochammad Ali Fauzi[2], Fitra Abdurrachman Bachtiar[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]amaliakusumaakaresti@gmail.com, [2]moch.ali.fauzi@ub.ac.id, [3]fitra.bachtiar@ub.ac.id

Abstrak

Bertambahnya jumlah situs berita *online* mengakibatkan terjadinya ledakan informasi dan terjadilah redundansi informasi. Redundansi informasi dapat menyebabkan pengguna internet menghabiskan bayak waktu apabila membaca bermacam berita dari situs berita *online* yang berbeda namun memiliki inti informasi yang sama. Dari permasalahan tersebut diperlukan sistem pencarian untuk memperudah mencari informasi, namun dengan hanya menggunakan mesin pencari pengguna internet masih harus membaca satu-persatu dari banyaknya informasi yang didapatkan. Oleh karena itu diperlukan pula sistem peringkasan untuk memudahkan pengguna internet menghindari mendapatkan informasi yang sama dari sumber yang berbeda. Pada penelitian ini dilakukan peringkasan multi-dokumen berbasis clustering pada sistem temu kembali berita *online* menggunakan metode *K-Means*. Proses sistem pencarian menggunakan metode *Cosine Similarity* dan pada peringkasan menggunakan metode pembobotan *Term Frequency-Inverse Sentence Frequency* (TF-ISF). Hasil menunjukan bahwa hasil yang optimal pada sistem pencarian sebesar *Recall* 71%, *Presicion* 65.82%, *F-Measure* 66.35% dan pada sistem peringkasan sebesar *Recall* 37.3%, *Presicion* 18%, *F-Measure* 19.2%

Kata kunci: *k-means clustering, multi-dokumen, peringkasan, sistem temu kembali informasi, TF-IDF, TF-ISF*

Abstract

The growing number of online news sites resulted in an explosion of information and information redundancy occurred. On this issue it takes the search engine to make it easier for users to find information, but users still have to read it one by one, therefore it needs also a summary system. Therefore a summary system is required to facilitate Internet users avoid getting the same information from different sources. In this study, multi-document clustering based on online news retrieval system using K-Means method. The process of searching system using Cosine Similarity method and on the summary using K-Means Clustering method. The results show that the optimum results in the recall system are Recall 71%, Presicion 65.82%, F-Measure 66.35% and on Recall system of Recall 37.3%, Presicion 18%, F-Measure 19.2%.

Keywords: *information retrieval, k-means clustering, multi-document, summarization, TF-IDF, TF-ISF*

1. PENDAHULUAN

Bertambahnya jumlah situs berita *online* mengakibatkan terjadinya ledakan informasi dan terjadilah redundansi informasi. Redundansi informasi dapat menyebabkan pengguna internet menghabiskan bayak waktu apabila membaca bermacam berita dari situs berita *online* yang berbeda namun memiliki inti informasi yang sama. Dari permasalahan tersebut diperlukan sistem pencarian untuk memperudah mencari informasi, namun dengan hanya menggunakan mesin pencari pengguna internet masih harus membaca satu-persatu dari banyaknya informasi yang didapatkan.

Sistem pencarian adalah konsep dasar dari sistem temu kembali informasi atau *Information Retrieval* (Maruhum, T., 2009). Salah satu penelitian sebelumnya pada sistem temu kembali informasi menerapkan *clustering* dokumen dengan metode *partitional*, yaitu algoritma *Bisecting K-Means* dan *Buckshot* metode *Hierarchical Agglomerative* dengan algoritma perhitungan similaritas kluster UPGMA dan *Complete Link* oleh Amir Hamzah (2009). Dari penelitian tersebut dapat membuktikan secara

signifikan meningkatkan kinerja *Information Retrieval* sebesar 12.3% dan 9.5% dibandingkan dengan metode linear berbasis *word-matching*. Ada pula penelitian oleh Fatkhul Amin dan Eddy Nurraharjo (2017) menggunakan metode *Cosine Similarity*, dari hasil penelitiannya sistem temu kembali informasi dokumen teks Berbahasa Indonesia memiliki nilai rata-rata *Recall* 0,02 dan rata-rata *Precision* 0,94. Sistem temu kembali informasi yang dibangun memiliki keunggulan memiliki hasil pencarian yang akurat serta dilengkapi dengan bobot dan letak dokumen pada *database*.

Peringkasan teks otomatis (*Automatic Text Summarization)* ialah bentuk singkat dari suatu teks dengan yang didapatkan menggunakan fungsi aplikasi yang dijalankan dan dioperasikan pada komputer. Sedangkan menurut Hovy (2001), ringkasan adalah teks yang dihasilkan dari sebuah teks atau banyak teks, yang mengandung isi informasi dari teks asli dan panjangnya tidak lebih dari setengah panjang teks aslinya.

Beberapa penelitian tentang peringkasan multi-dokumen berita, yang salah satunya dilakukan oleh Muztahid (2015). penelitian yang dilakukan memberikan output ringkasan dokumen skripsi dengan hasilkan rata-rata akurasi 58.51%, *Recall* 22.06%, Precision 43.84%, dan *F-Measure* 27.88%.

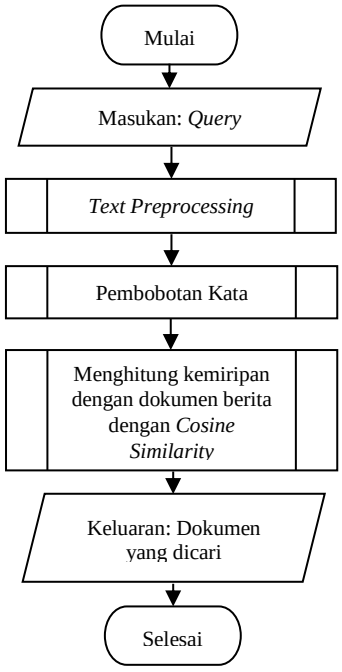
Dengan metode yang sama, Yudi Wibisono dan Masayu Leylia Kohdra (2006) melakukan pengelompokan berita berbahasa Indonesia yang mengalami peningkatan dalam jumlah besar disetiap harinya. Dalam penelitiannya melakukan pengujian hasil dengan menguji coba penggunaan *Stemming*, untuk melihat pengaruhnya terhadap kualitas cluster yang dihasilkan. Penelitian ini menunjukan bahwa penggunaan log-tf.idf tanpa menggunaan Stemming menghasilkan kualitas cluster terbaik dengan rata-rata purity 0.568. Namun kualitas cluster masih rendah dan perlu dilakukan penelitian lanjutan.

Berdasarkan penjelasan sebelumnya, peneliti ingin melakukan penelitian dengan menggabung sistem pencarian dengan peringkasan dengan metode *K-Means Clustering* tanpa melakukan proses *Stemming* pada tahap *text preprocessing*.

2. METODE USULAN

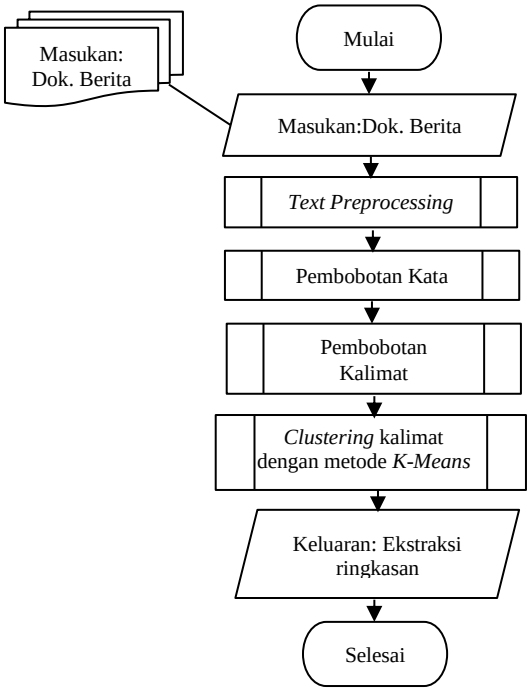
Peringkasan multi-dokumen berbasis *Clustering* pada sistem temu kembali berita *online* menggunakan metode *K-Means* merupakan sistem yang dikembangkan untuk mepermudah pengguna internet dalam menerima berita secara ringkas tanpa harus membuang waktunya dengan membaca banyak berita dengan topik yang sama dan menghasilkan peringkasan yang efektif dengan efisien.

Sistem pada penelitian ini terdapat dua tahap pemrosesan, yaitu pencarian dan peringkasan. Tahap pencarian merupakan tahap dimana sebuah kata kunci atau *query* yang diberikan oleh pengguna sistem dimasukan ke dalam sistem dan kemudian sistem akan memprosesnya yang nantinya sistem akan memberikan keluaran ringkasan berita sesuai dengan *query* yang diberikan. Ada beberapa subproses yang harus dilakukan pada tahap ini, yaitu *Text Preprocessing*, pembobotan kata, *Cosine Similarity* dan *K-Means clustering*. Alur tahap pencarian akan dijelaskan pada Gambar 1.



Gambar 1. Diagram Alir Sistem Pencarian

Sedangkan untuk tahap peringkasan adalah tahap dimana dokumen yang memiliki kemiripan dengan *query* yang diberikan oleh pengguna, nantinya akan diringkas dan ditampilkan sebagai hasil keluaran sistem. Ada beberapa subproses pula yang harus dilakukan pada tahap ini, yaitu *Text Preprocessing*, pembobotan kata, pembobotan kalimat, *Cosine Similarity* dan *K-Means clustering*. Alur tahap peringkasan ditunjukkan pada Gambar 2.



Gambar 2. Diagram Alir Sistem Peringkasan

2.1 Data

Data berita yang digunakan pada penelitian ini diambil dari web berita *online* yaitu; Tribunnews.com, Detik.com, Liputan6.com, dan Kompas.com. Data yang digunakan berupa berita dengan topik 'Badan Siber Nasional' berjumlah 200 berita dan diuji dengan variasi *query* yang telah di tentukan.

2.2 Pembobotan Kata (*Term Weighting*)

Setiap kata atau term memiliki nilai atau bobotnya masing-masing, pada penelitian ini menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) sebagai metode *term weighting*. *Term frequency* (TF) adalah jumlah munculnya suatu term dalam suatu dokumen. Manning et al. (2008) menjelaskan bahwa metode *Inverse Document Frequency* (IDF) baik digunakan untuk peringkasan teks. IDF merupahan operasi pembagian jumlah dokumen dengan frekuensi dokumen yang membuat suatu term. sedangkan TF-IDF adalah hasil perkalian nilai TF dengan IDF untuk sebuah term dalam dokumen. Persamaan *IDF* dan TF-IDF dapat dilihat pada persamaan (1) dan (2) dibawah ini.

$$IDF_t = log(\frac{N}{DF_t}) \quad (1)$$

$$TF - IDF_t = TF * log(\frac{N}{DF_t}) \quad (2)$$

Keterangan (1) dan (2):

N : jumlah dokumen yang berisi *term* (t)
DF_t : jumlah kemunculan kata (*term*) terhadap N

2.3 Pembobotan Kalimat (*Sentence Scoring*)

Term Frequency – Inverse Sentence Frequency (TF-ISF) adalah salah satu fase penting pada peringkasan berita adalah pembobotan kalimat. Perhitungan ISF dan TF-ISF dapat dirumuskan pada persamaan (3) dan (4) sebagai berikut (Muztahid, 2015):

$$ISF = \log(\frac{N}{SF_t}) \quad (3)$$

$$TF\ ISF_{t,s} = TF_{t,s} \times ISF_t \quad (4)$$

Keterangan (3) dan (4):
N : jumlah banyaknya kalimat dalam dokumen
SF_t : banyaknya kalimat yang mengandung kata t

2.4 Cosine Similarity

Cosine Similarity merupakan metode yang digunakan untuk mengukur kemiripan antar dua dokumen atau lebih, serta untuk menghitung nilai cosinus sudut antara dua *vector* (Cahyanti, Saptono, & Sari, 2015). Perhitungan *similarity* dalam membandingkan parameter tersebut dapat dihitung dengan persamaan (5) sebagai berikut:

$$similarity = cos\theta = \frac{A.B}{\|A\|.\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)} \times \sqrt{\sum_{i=1}^{n}(B_i)}} \quad (5)$$

Keterangan:
A = bobot data yang dibandingkan
B = bobot data pembanding
$\|A\|$= panjang data yang dibandingkan
$\|B\|$= panjang data pembanding

2.5 K-Means Clustering

K-Means Clustering adalah metode populer yang digunakan untuk mendapatkan dekripsi dari sekumpulan data dengan cara mengungkapkan kecenderungan setiap individu data untuk berkelompok dengan individu-individu data lainnya. Kecenderungan pengelompokan tersebut didasarkan pada kemiripan karakteristik individu-individu data yang ada (Prilianti & Wijaya, 2014).

Algoritma *K-Means* selanjutnya akan melakukan pengulangan langkah-langkah berikut sampai terjadi kestabilan (tidak ada obyek yang dapat dipindahkan) (Deshpande & L, 2013):

1. Menentukan koordinat titik tengah setiap *cluster*,
2. Menentukan jarak setiap obyek terhadap koordinat titik tengah,
3. Mengelompokkan obyek-obyek tersebut berdasarkan pada jarak minimumnya.

Rumus yang digunakan untuk menghitung jarak data dengan centroid adalah rumus *Euclidean Distance*. Adapun rumus *Euclidean Distance* dapat dilihat pada persamaan (6) berikut:

$$d(x_i c_j) = \sqrt{\sum_{j=1}^{n}(x_{ik} - c_{jk})^2} \qquad (6)$$

Keterangan (6):
d = jarak data dengan *centroid*
j = banyaknya data
k = dimensi
c = *centroid*
x = data

3. HASIL DAN PEMBAHASAN

Pada penelitian ini dilakukan dua macam penelitian. Pengujian yang pertama pada penelitian ini dilakukan dengan perhitungan *Precison*, *Recall*, *F-Measure* yang diambil dari hasil penilaian oleh pakar. Pengujian yang dilakukan oleh pakar sendiri dengan menggunakan 200 dokumen artikel berita *online* dan 10 macam variasi *query* yang dimasukan, variasi *query* ada pada Tabel 1. Hasil pengujiannya akan dijelaskan dengan dua jenis pengujian, yaitu pengujian terhadap sistem temu kembali berita dan pengujian terhadap peringkasan. Kedua jenis pengujian tersebut akan diberlakukan 3 macam skenario yang berbeda. Masing-masing skenario diambil nilai rata-ratanya dan kemudian dibandingkan untuk melihat mana hasil yang paling optimal.

Salah satu fungsi sistem dari penelitian ini adalah kemudahan dalam pencarian informasi. Maka pada pengujian yang kedua dilakukanlah pengujian yang duji cobakan sistem secara langsung oleh pengguna umum.

Tabel 1. Daftar *Query*

	Variasi *Query*:	Keterangan
1.	Siber	1 kata
2.	Keamanan	1 kata
3.	Joko Widodo	2 kata
4.	Ketua BSSN	2 kata
5.	Badan Siber Nasional	3 kata
6.	Jokowi Resmikan BSSN	3 kata
7.	Badan Siber Sandi Negara	4 kata
8.	Tugas Badan Siber Nasional	4 kata
9.	Badan Siber Nasional Gagal Dibentuk	5 kata
10.	Peran BSSN Jaga Keamanan Siber	5 kata

3.1 Hasil Pengujian oleh Pakar

1. Sistem Temu Kembali Berita

Pengujian sistem temu kembali berita, dari 10 *query* yang dimasukan dilakukan 3 macam skenario pengujian. Macam skenario yang divariasi adalah jumlah dokumen berita dengan ranking tertinggi hasil *Cosine Similarity*, macam yang digunakan yaitu 4 dokumen tertinggi, 7 dokumen tertinggi dan 10 dokumen tertinggi. Dari 3 skenario tersebut dihasilkan nilai optimal pada scenario 3, yaitu dengan 10 dokumen tertinggi, dengan perbandingan nilai yang dapat dilihat pada Tabel 2.

Tabel 2. Perbandingan hasil 3 skenario Pengujian Sistem Temu Kembali Berita

Skenario	Jumlah Dokumen	Precison	*Recall*	*F-Measure*
1	4 dok.	70.00%	26.16%	37.03%
2	7 dok.	72.86%	53.87%	59.75%
3	10 dok.	71.00%	65.82%	66.35%

Pada skenario 3 nilai rata-rata *F-Measure* yang dihasilkan lebih besar daripada skenario 1 dan skenario 2, sehingga dapat disimpulkan bahwa hasil pengujian sistem temu kembali berita dengan skenario 3 lebih unggul dari pengujian dengan skenario 1 dan skenario 2. sistem temu kembali berita dengan skenario 3 lebih unggul daripada skenario 1 dan skenario 2 karena pengujian sistem temu kembali berita dengan skenario 3 memiliki sumber data yang digunakan lebih banyak dari sumber data yang digunakan pada skenario 1 dan skenario 2. Semakin banyak jumlah data yang digunakan, semakin optimal nilai *F-Measure* yang diperoleh.

Pada variasi *query* yang digunakan, nilai rata-rata *F-Measure* yang diperoleh hasil yang maksimal di skenario 3 pada *query* 3, *query* 4 dan *query* 7. Sedangkan nilai *F-Measure* minimal pada *query* 10. Melihat hasil pengujian yang tidak memiliki kenaikan nilai yang signifikan pada jumlah kata pada *query* yang diujikan di ketiga skenario, makan dapat disimpulkan pula bahwa nilai optimal pada *query* yang dimasukan tergantung pada jumlah informasi yang tersedia pada data yang ada. Semakin banyak jumlah data yang digunakan,

semakin besar kemungkinan informasi yang tersedia akan lebih banyak pula.

2. Peringkasan

Pengujian peringkasan, dari 10 *query* yang dimasukan dilakukan 3 macam skenario pengujian. Macam skenario yang divariasi adalah nilai *cluster* yang digunakan untuk menentukan keluaran atau jumlah kalimat peringkasan sistem, macam yang digunakan yaitu 4 kalimat, 7 kalimat dan 10 kalimat. Dari 3 skenario tersebut dihasilkan nilai optimal pada skenario 3, yaitu 10 kalimat. Hasil ditungjukan pada Tabel 3.

Tabel 3. Perbandingan hasil 3 skenario Pengujian Peringkasan

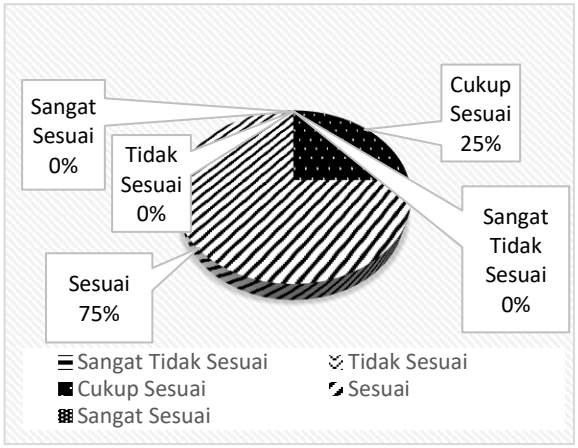
Skenario	Jumlah *Cluster*	Precison	*Recall*	*F-Measure*
1	4 *cluster*	40.33%	14.06%	13.92%
2	7 *cluster*	42.48%	16.49%	17.50%
3	10 *cluster*	37.33%	18.01%	19.18%

Hasil yang didapatkan dari ketiga skenario di pengujian peringkasan yang ditampilkan pada tabel 5 terbilang rendah dari pengujian sistem temu kembali berita. Terutama di nilai precisionnya, nilai precision sendiri bergantung pada hasil yang dihasilkan oleh pakar. Pakar yang digunakan dalam penelitian ini orang yang berkecimpung di dunia jurnalistik, dan setelah mengkaji antara hasil sistem dan hasil pakar dapat disimpulkan bahwa hasil *clustering* peringkasan oleh pakar untuk hasil ringkasan yang dihasilkan, pakar memperhatikan susuan antar kalimatnya kalimatnya dan mempenhatikan kesesuaian topik antar *query* yang dimasukan dengan informasi yang tersedia pada data. Sedangkan sistem dalam hasil peringkasannya hanya bergantung pada nilai *Euclidean Distance*, atau kesesuaian nilai jarak kedekatan antar centroidnya. Sehingga ini menjadi benyebab mengapa hasil pengujian pada peringkasan memiliki hasil yang sangat rendah. Oleh karena itu diperlukan pengujian lanjutan yang akan dibahas pada subbab 3.2.

3.2 Hasil Pengujian oleh Pengguna Umum

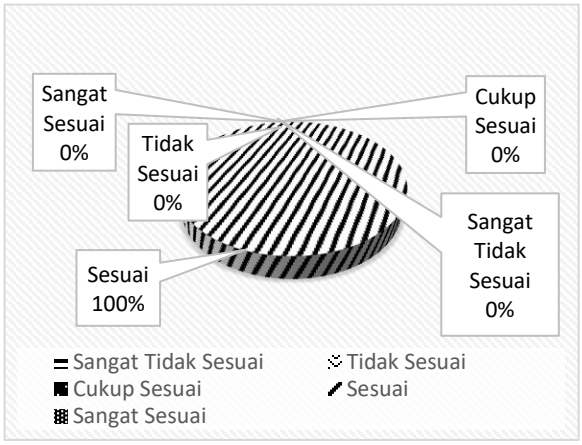
Pengujian yang kedua ini dilakukan guna untuk penyeimbang hasil pengujian dari pengujian yang pertama, yang mana dilakukan oleh seorang pakar jurnalis yang memiliki penilaian khusus tentang bagaimana sebuah ringkasan harus berkesesuaian dengan substansi dari *query* yang diberikan. Kegunaan pengujian kedua ini pula untuk menilai tingkat kepuasan dari sistem yang telah dibuat. Terdapat dua jenis penilaian kepuasan, yaitu kemudahan dalam pencarian informasi dan kepusan terhadap penggunaan sistem dalam penelitian ini. Pengujian oleh pengguna umum ini dilakukan oleh 5 responden dan 5 range penilaian untuk sistem yang telah dibuat dan diuji cobakan, yaitu 1. Sangat tidak setuju, 2. Tidak setuju, 3. Cukup setuju, 4. Setuju, 5. Sangat Setuju. Hasil pengujian untuk kemudahan pencarian informasi dapat dilihat pada Gambar 3 dan hasil kepuasan pengguna dapat dilihat pada Gambar 4.

1. Kemudahan Pencarian Informasi



Gambar 3. Hasil Penilaian Kemudahan Sistem dalam Pencarian Informasi

2. Kepuasan Penggunaan Sistem



Gambar 4. Hasil Penilaian Kepuasan Responden dengan Hasil Ringkasan

Dari hasil yang didapatkan pada pengujian kemudahan dan kepuasan oleh pengguna umum, dapat disimpulkan bahwa sistem pencarian yang dibuat sudah sesuai dengan tujuan penelitian untuk memudahkan pengguna internet dalam mencari informasi.

4. KESIMPULAN

Berdasarkan seluruh tahapan yang telah dilakukan dalam penelitian, berikut beberapa point sebagaimana hasil yang diperoleh:

1. Pada pengujian sistem temu kembali berita dilakukan pengujian dengan 3 skenario diperoleh nilai paling optimal pada skenario 3 dengan nilai rata-rata *Recall* sebesar 71%, *Precision* sebesar 65.82%, dan *F-Measure* 66.35%. Dari hasil yang didapatkan, sistem temu kembali berita menggunakan pembobotan TF-IDF dan metode *Cosine Similarity* menunjukan bahwa hasil yang didapatkan akan mencapai nilai optimal dengan penggunaan sumber atau jumlah data yang banyak.
2. Pada pengujian sistem peringkasan dilakukan pengujian dengan 3 skenario diperoleh nilai pailng optimal pada skenario 3 dengan nilai rata-rata *Recall* sebesar 37.3%, *Precision* sebesar 18%, dan *F-Measure* 19.2%. Dari hasil yang didapatkan, sistem peringkasan menggunakan pembobotan TF-ISF metode *K-Means* menunjukan bahwa hasil yang didapatkan akan mencapai nilai optimal dengan penggunaan cluster atau jumlah kalimat ringkasan yang lebih banyak.
3. Dari hasil yang didapatkan pada pengujian kemudahan dan kepuasan oleh pengguna umum sebagai responden penelitian ini, 75% suara mengatakan sesuai dan 25% suara mengatakan cukup sesuai dalam memudahkan dalam mencari informasi. Dan didapatkan 100% suara yang mengatakan bahwa responden puas dengan informasi dari hasil ringkasan yang didapatkan sistem. Dari pengujian yang dilakukan oleh responden ini dapat disimpulkan bahwa sistem pencarian yang dibuat sudah sesuai dengan tujuan penelitian untuk memudahkan pengguna internet dalam mencari dan mendapatkan informasi secara efektif dan efisien.

Penelitian selanjutnya dapat menggunakan jumlah data yang lebih banyak lagi dan variasi *query* memperhatikan ketersediaan informasi yang ada pada data yang digunakan. Menggunakan jenis peringkasan abstraktif dengan fitur semantik sehingga hasil ringkasan dapat berkesinambungan antara satu kalimat dengan lainnya dan yang dapat memperhatikan kesesuaian topik atau substansi antara *query* yang dimasukan dengan informasi dari ringkasan yang didapatkan. Melakukan *Stemming* sehigga dapat lebih membantu menyaring kata kunci yang dimasukan.

5. DAFTAR PUSTAKA

Aditya, C. S., Fatichah, C., & Purwitasari, D. (2016). Ekstraksi Trending Issue dengan Pendekatan Distribusi Kata Pada Pembobotan Term untuk Peringkasan Multi-Dokumen Berita. *JUTI*, 180-189.

Amin, F. & Nurraharjo, E., 2017. Rekayasa Sistem Temu Kembali Informasi Dokumen Teks Berbahasa Jawa Metode Cosine Similarity Dan Rule Base Stemming Bahasa Jawa. Prosiding SINTAK, pp. 41-48.

Aristoteles. (2013). Penerapan Algoritma Genetika pada Peringkasan Teks Dokumen Bahasa Indonesia. *Prosiding Semirata FMIPA Universitas Lampung*, 29-33.

Azhar, R., Machmud, M., Hartanto, H. A., Arifin, A. Z., & Purwitasari, D. (2016). Pembobotan Kata Berdasarkan Klaster Pada Optimisasi Coverage, Diversity Dan Coherence Untuk Peringkasan Multi Dokumen. *Jurnal Ilmiah Teknologi Informasi Terapan, II*(3), 170-178.

B, Z., & E, W. (2011). Analisis Fitur Kalimat untuk Peringkas Teks Otomatis pada Bahasa Indonesia. *IJCCS*, Vol.5 No.2.

Cahyanti, A. F., Saptono, R., & Sari, W. S. (2015). Penentuan Model Terbaik pada Metode Naive Bayes Classifier dalam Menentukan Status Gizi Balita dengan Mempertimbangkan Independensi Parameter. *JURNAL ITSMART*, Vol. 4 No. 1.

Deshpande, A. R., & L, L. M. (2013). Text Summarization using *Clustering* Technique. *International Journal of Engineering Trends and Technology (IJETT)*, Volume4 Issue8.

Hamzah, A., 2009. Temu Kembali Informasi Berbasis Kluster Untuk Sistem Temu Kembali Informasi Teks Bahasa Indonesia. *Jurnal Teknologi*, Volume Volume 2 Nomor 1, pp. 1-7.

Hassel, M. (2004). Summaries and the Process of Summarization from Evaluation of

Automatic Text Summarization. *A practical Implementation, Liccentiate Thesis, KTH NADA*.

Hovy E. H. 2001. Automated Text Summarization. In R. Mitkov (Ed.). Hanbook of computation linguistics. Oxford University Press.

Kurniawan, B., Effendi, S., & Sitompul, O. S. (2012). Klasifikasi Konten Berita Dengan Metode Text Mining. *JURNAL DUNIA TEKNOLOGI INFORMASI Vol. 1, No. 1*, 14-19.

Lazuardy, M. M. (2016). Comparison of Vector Space Model and Support Vector Machine For Text Summarization Against Indonesian Language Articles. *JBPTUNIKOMPP*.

Luthfiarta, A., Zeniarja, J., & Salam, A. (2013). *Algoritma Latent Semantic Analysis.* SEMANTIK.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *An Introduction to Information Retrieval.* Cambridge (UK): Cambridge Univercity Press.

Maruhum, T.(2009). Analisis Dan Implementasi Query Expansion Pada Information Retrieval Berdasarkan Penarikan Kesimpulan Dengan Fuzzy Rules. In Witten, Ian H., Moffat, Alistair, Bell,Timothy C. Managing Gigabytes: Compressing and Indexing Documents and Images, second edition. Morgan Kaufmann Publishers, Academic Press.

Mustaqhfiri, M., Abidin, Z., & Kusumawati, R . (2011). Peringkasan Teks Otomatis Berita Berbahasa Indonesia Menggunakan Metode Maximum Marginal Relevance. *MATICS*, 4(4): 134-147.

Muztahid, M. R. (2015). *Peringkasan Dokumen Bahasa Indonesia Menggunakan Metode K-Means.* Bogor: Institut Pertanian Bogor.

Prilianti, K. R., & Wijaya, H. (2014). Aplikasi Text Mining untuk Automasi Penentuan Tren Topik Skripsi dengan Metode *K-Means Clustering*. *Jurnal Cybermatika*, Vol. 2 No. 1.

Triawati, C. (2009). *Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi Dokumen Bahasa Indonesia. Skripsi.* Bandung: Institut Teknologi Telkom Bandung.

Wibisono, Y., & Khodra, M. L. (2006). *Clustering* Berita Berbahasa Indonesia.

Yang, J., Qu, Z., & Liu, Z. (2014). Improved feature selection method considering the imbalance problem in text categorization. *Hindawi Publishing Corporation the Scientific World Journal*.