INTEGRASI PERINGKAS DOKUMEN OTOMATIS SEBAGAI FEATURE REDUCTION PADA CLUSTERING DOKUMEN

Abu Salam[1], Catur Supriyanto[2], Amiq Fahmi[3]
[1,2]Magister Teknik Informatika, Univ. Dian Nuswantoro
Email: masaboe@yahoo.com
[3] Manajemen Informatika, Univ. Dian Nuswantoro
Email: amfa_dns@yahoo.com

ABSTRAK

Clustering dokumen merupakan proses pengelompokan dokumen yang memiliki kesamaan topik, clustering dokumen memudahkan pengguna menemukan dokumen yang diinginkan. Dalam proses clustering dokumen, dokumen direpresentasikan menggunakan Vector Space Model (VSM). Masalah dalam VSM adalah matrik term-dokumen biasanya sangat jarang (banyak mengandung angka 0 dalam term-dokumen matrik) dan juga mempunyai dimensi tinggi, sehingga masalah-masalah ini dapat mengurangi kinerja clustering dokumen. Oleh karena itu diperlukan suatu metode untuk bisa mengurangi dimensi term-dokumen dan menghilangkan term yang bernilai 0 tersebut sehingga dapat meningkatkan kinerja proses clustering. Dalam penelitian ini diusulkan model peringkas dokumen otomatis sebagai feature reduction pada proses clustering dokumen.

Tujuan dari penelitian ini adalah untuk meningkatkan akurasi dari clustering dokumen dengan mengintegrasikan peringkas dokumen otomatis sebagai feature reduction. Ada beberapa tahapan clustering dalam penelitian ini, yaitu preprocessing, peringkas dokumen otomatis, pembobotan kata, feature selection, feature transformation dan algoritma clustering. Tahap Preprocessing yang digunakan dalam penelitian ini adalah tokenization, stopword, stemming dan pemenggalan kalimat. Proses peringkas dokumen otomatis ditujukan untuk penyeleksian kalimat agar didapatkan ringkasan teks yang diperoleh dengan menyajikan kembali bagian tulisan yang dianggap topik utama tulisan dengan bentuk yang lebih disederhanakan baru kemudian selanjutnya dilakukan proses pembobotan kata, feature selection, feature transformation dan clustering.

Hasil penelitian menunjukkan bahwa integrasi peringkas dokumen otomatis sebagai feature reduction dapat meningkatkan kinerja clustering dokumen sampai dengan 91,7 %, mengalami peningkatan dari tingkat akurasi 89,6 % untuk proses feature reduction tanpa menggunakan peringkas dokumen otomatis. Kemudian pengaruh Integrasi peringkas dokumen otomatis sebagai feature reduction untuk waktu komputasi yang dibutuhkan adalah pada % feature selection yang semakin kecil integrasi peringkas dokumen otomatis sebagai feature reduction membutuhkan tambahan waktu komputasi tersendiri, akan tetapi pada proporsi feature selection yang semakin besar, % peringkas dokumen otomatis dapat menurunkan waktu komputasi yang digunakan.

.***Kata kunci**:* Text mining; Clustering Dokumen; Peringkas Dokumen Otomatis.

1. Latar Belakang

Clustering dokumen adalah proses pengelompokan dokumen yang memiliki kesamaan topik, c*lustering* dokumen memudahkan pengguna menemukan dokumen yang diinginkan [1]. Dengan semakin banyaknya volume dokumen yang ada, dapat menyebabkan suatu permasalahan pada clustering dokumen yaitu besarnya matrik term-dokumen yang bisa menyebabkan proses kerja clustering dokumen tidak optimal. Hal ini bisa terjadi karena adanya data yang tidak relevan dan redundan. Oleh karena itu diperlukan suatu metode untuk bisa mengurangi dimensi dokumen tersebut sehingga bisa meningkatkan kinerja proses clustering dokumen tanpa mengurangi tingkat akurasi hasil clustering [2] [3].

Ringkasan dokumen dapat diartikan sebagai proses dari pembuatan intisari informasi terpenting dari sumber untuk menghasilkan versi yang lebih ringkas, terdapat dua tipe pembuatan suatu ringkasan yang mengambil bagian terpenting dari teks aslinya yaitu abstak dan ekstrak. Abstrak menghasilkan sebuah interprestasi terhadap teks aslinya, dimana sebuah kalimat akan ditransformasikan menjadi kalimat yang lebih singkat, sedangkan ekstraksi merupakan ringkasan teks yang diperoleh dengan menyajikan kembali bagian tulisan yang dianggap topik utama tulisan dengan bentuk yang lebih disederhanakan [4]. Dalam penelitian ini akan digunakan fitur ringkasan ekstrak sebagai model peringkas dokumen otomatis. Sebagai pembuktian akan dibandingkan akurasi proses clustering menggunakan feature reduction standar dengan proses clustering yang menggunakan peringkas dokumen otomatis sebagai feature reduction.

2. Dasar Teori

2.1 Clustering Dokumen

Algoritma yang umum digunakan pada proses clustering dokumen dan juga yang akan digunakan dalam penelitian adalah algoritma *K-means*, Dasar algoritma *K-means* dapat disusun menjadi 4 tahap sebagai berikut:

1. Inisialisasi titik pusat Cluster
2. Masukan setiap dokumen ke cluster yang paling cocok berdasarkan ukuran kedekatan dengan centroid / titik tengah cluster.
3. Setelah semua dokumen masuk ke cluster. Hitung ulang centroid cluster berdasarkan dokumen yang berada di dalam cluster tersebut.
4. Jika centroid tidak berubah (dengan treshold tertentu) maka stop. Jika tidak, kembali ke langkah 2.

$$Sim(d_x,d_y) = \frac{\sum_{k=1}^{n} x_k \times y_k}{\sqrt{\sum_{k=1}^{n} x_k^2} \times \sqrt{\sum_{k=1}^{n} y_k^2}} \quad (1)$$

2.2 Preprocessing

Preprocessing merupakan tahapan untuk mengubah struktur isi dari suatu dokumen kedalam format yang sesuai berupa kumpulan term atau kata untuk diproses kedalam algoritma clustering [4], dalam penelitian ini digunakan 4 tahapan preprocessing yaitu: Tokenization, Stopword, Stemming dan Sentence splitting.

2.3 Document Representation Vector Space Model

VSM mengubah koleksi dokumen kedalam matrik *term-document* [2]. Pada gambar 4.3.1. Dimana *d* adalah dokumen dan *w* adalah bobot atau nilai untuk setiap term.

$$A_{mxn} = \begin{array}{c} \begin{matrix} d1 & d2 & dn \\ \downarrow & \downarrow & \downarrow \end{matrix} \\ \begin{bmatrix} \omega_{11} & \omega_{11} & \cdots \omega_{11} \\ \omega_{11} & \omega_{11} & \cdots \omega_{11} \\ \vdots & & \vdots \\ \omega_{m1} & \omega_{m2} & \cdots \omega_{mn} \end{bmatrix} \end{array} \begin{array}{l} \leftarrow t1 \\ \leftarrow t2 \\ \\ \leftarrow tm \end{array}$$

Gambar 1. Matrik Term-dokumen

2.4 Term Weighting (TFIDF)

TF adalah banyaknya kemunculan suatu *term* dalam suatu dokumen, IDF adalah perhitungan logaritma antara pembagian jumlah total dokumen dengan cacah dokumen yang mengandung suatu *term*, dan TFIDF adalah perkalian antara TF dengan *IDF*. Dalam penelitian ini digunakan TFIDF sebagai metode *term weighting*.

$$IDF = log \frac{D}{DF} \quad (2)$$

$$TFIDF(t) = TF * log \frac{D}{DF} \quad (3)$$

2.5 Similiarity Measure

Pada Vector Space Model Dokumen direpresentasikan dalam bentuk $d = \{w_1, w_2, w_3, \ldots, w_n\}$ dimana *d* adalah dokumen dan *w* adalah nilai bobot setiap term dalam dokumen. Dalam penelitian ini untuk menghitung persamaan antar dokumen akan mengukur jarak antar 2 dokumen d_i dan d_j dengan menggunakan rumus *cosines similiarity*.

$$similiarity\ (d_i, d_j) = cosines\ \theta = \frac{\vec{di} \cdot \vec{dj}}{||di|| \cdot ||dj||} \quad (4)$$

2.6 Teknik Dimension Reduction

2.6.1 Feature Selection

Disebutkan bahwa hasil dari clustering teks mempunyai ketergantungan dengan kesamaan dokumen. sehingga, kontribusi dari sebuah term dapat diartikan sebagai kontribusi terhadap kesamaan dokumen.

$$TC(t) = \sum_{i,j \cap 1 \neq j} f(t, di).f(t, dj) \quad (5)$$

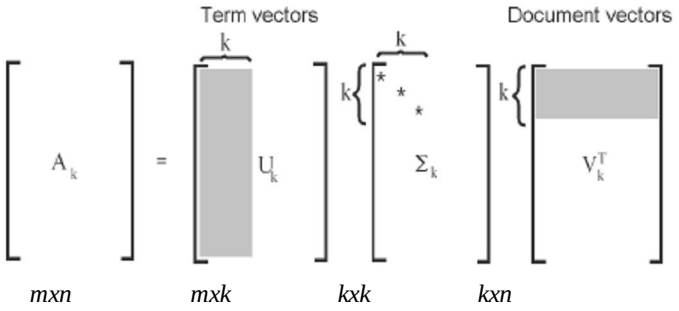
Di mana, $f(t,d)$ merupakan bobot *tf*idf* dari term *t* di dokumen *d*.

2.6.2 Singular Value Decomposition

Latent Semantic Indexing (LSI) melalui metode Singular Value Decomposition (SVD) mengurai matrik term-document menjadi 3 matrik *U*, *S* dan *V* yang memiliki dimensi lebih kecil.

$$A = USV^T \quad (6)$$

Dimana *U* merupakan matrik term yang berdimensi *m* x *k*, *S* adalah matrik diagonal yang berisi eigen value berdimensi kxk dan V^T adalah matrik dokumen yang memiliki dimensi *k* x *n*.



Gambar 2. Dekomposisi truncated SVD.

Truncated SVD menggunakan pendekatan rank-k untuk mengurangi SVD [5], Dalam penelitian ini menggunakan peringkat-k pembulatan nilai akar dari jumlah 150 dokumen yang diproses, yaitu pembulatan dari $\sqrt{150} = 12$.

2.7 Feature-Based Automatic Summarization

6 fitur tahapan yang digunakan dalam penelitian ini adalah sebagai berikut:

No	Tahapan	
1	Sentence length	$score(S) = \frac{No.Word\ occuring\ in\ S}{No.Word\ occuring\ in\ longest\ sentence}$
2	Term weight	$score(S) = \frac{Sum\ of\ TF-IFS\ in\ S}{Max(Sum\ of\ TF-IFS)}$
3	Sentence position	$score(s) = 1\ for\ first\ and\ last\ sentence.$ $0\ for\ other\ sentences.$
4	Sentence to sentence similiarity	$score(S) = \frac{Sum\ of\ Sentence\ Similiarity\ in\ S}{Max(Sum\ of\ sentence\ Similiarity)}$
5	Thematic word	$score(s) = \frac{No.Thematic\ word\ in\ S}{Length\ (S)}$
6	Numerical data	$score(s) = \frac{No.Numerical\ data\ in\ S}{Length\ (S)}$

Gambar 3. Enam Tahap Peringkas Dokumen Otomatis.

2.8 Evaluation Measure

Recall dan precision kategori *i* dalam cluster *j* diperoleh dari persamaan berikut :

$$Recall\ (i,j) = \frac{nij}{ni} \quad (7)$$

$$Precision\ (i,j) = \frac{nij}{nj} \quad (8)$$

Dinama n_{ij} merupakan jumlah dokumen kategori *i* dalam cluster *j*, n_i adalah jumlah dokumen dalam kategori *i* dan n_j merupakan jumlah dokumen dalam cluster *j*. kemudian untuk menghitung F-measure yang digunakan adalah persamaan berikut:

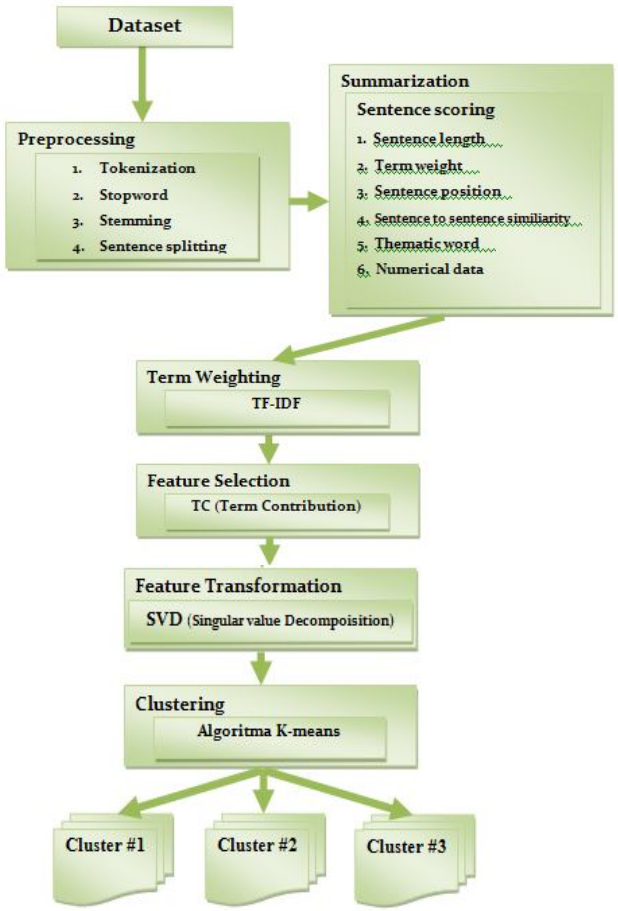
$$F(i,j) = \frac{2*(Precision*Recall)}{(Precision+Rscall)} \quad (9)$$

Secara keseluruhan, rata-rata dari F-measure dapat dihitung dengan persamaan berikut:

 ISBN 979 - 26 - 0255 - 0

$$F = \sum_i \frac{ni}{n} max_{j=1,..,k} F(i,j) \quad (10)$$

3. Metode yang Diusulkan



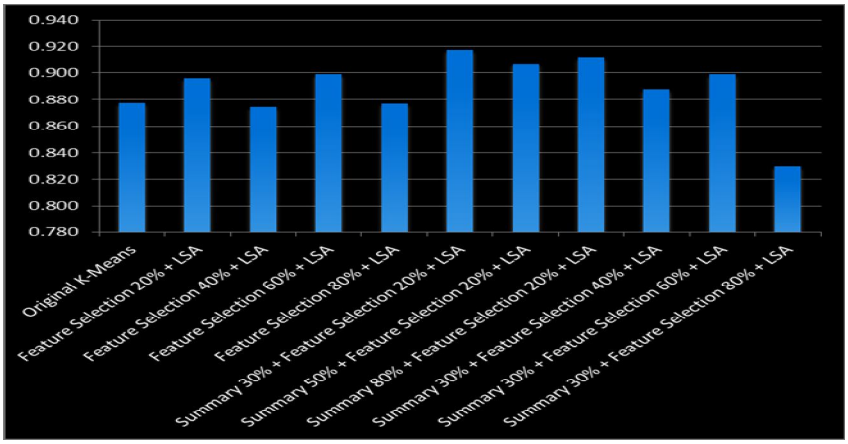
Gambar 4. Model yang diusulkan.

4. HASIL DAN PEMBAHASAN

4.1 Akurasi

Dari hasil penelitian yang dilakukan dapat dibuktikan bahwa integrasi peringkas dokumen otomatis sebagai feature reduction dapat meningkatkan akurasi hasil clustering. Tingkat akurasi menggunakan peringkas dokumen otomatis yang diintegrasikan sebagai feature reduction mencapai 91,7 % yang diperoleh pada tingkat peringkas dokumen otomatis 30 % dan feature selection 20 %, dibandingkan dengan feature selection 20 % tanpa menggunakan peringkas dokumen otomatis yang hanya mencapai tingkat akurasi 89,6 %. Dari gambar 5 juga dapat dilihat peningkatan akurasi untuk % feature selection yang lain, akan tetapi pada proporsi 80 % feature selection integrasi peringkas dokumen otomatis mengalami penurunan tingkat akurasi.

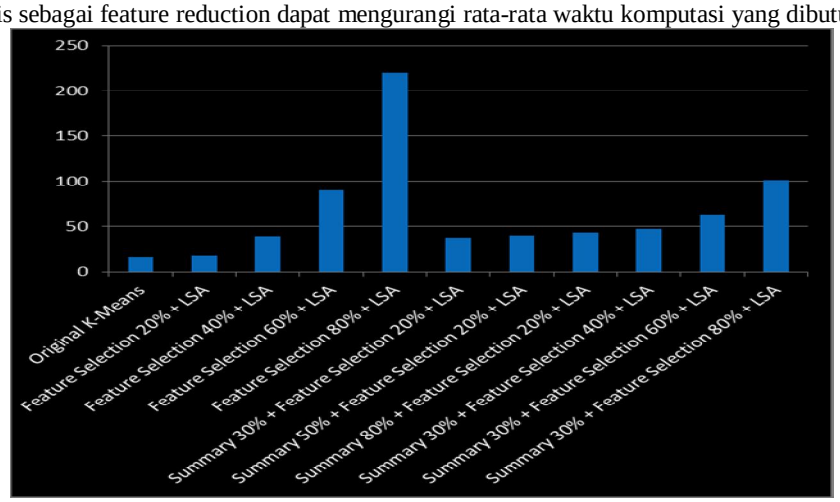
 ISBN 979 - 26 - 0255 - 0



Gambar 5. Hasil kinerja proses clustering dokumen.

4.2 Waktu

Waktu rata-rata yang diambil diukur mulai dari proses preprocessing sampai dengan hasil clustering diperoleh. Gambar 6 menunjukan bahwa pada % feature selection yang semakin kecil feature reduction yang diintegrasikan dengan peringkas dokumen otomatis membutuhkan tambahan waktu komputasi tersendiri, dari percobaan yang dilakukan untuk 20% feature selection terdapat peningkatan waktu komputasi dari percobaan clustering tanpa peringkas dokumen otomatis, menggunakan peringkas dokumen otomatis dengan proporsi 30%, 50% dan 80%. Akan tetapi pada proporsi feature selection yang semakin besar, % peringkas dokumen otomatis dapat menurunkan waktu komputasi yang ada, pada percobaan 60% dan 80% feature selection dapat dilihat bahwa intergrasi peringkas dokumen otomatis sebagai feature reduction dapat mengurangi rata-rata waktu komputasi yang dibutuhkan.



Gambar 6. Waktu proses clustering dokumen.

5. KESIMPULAN

Integrasi peringkas dokumen otomatis sebagai feature reduction pada proses clustering dokumen dapat meningkatkan tingkat akurasi hasil clustering. Hasil penelitian menunjukkan bahwa integrasi peringkas dokumen otomatis sebagai feature reduction tersebut dapat meningkatkan kinerja clustering dokumen sampai dengan 91,7 %, mengalami peningkatan dari tingkat akurasi 89,6 % untuk proses feature reduction standar tanpa menggunakan peringkas dokumen otomatis dan 87,73 % tingkat akurasi clustering standar. Pengaruh Integrasi peringkas dokumen otomatis sebagai feature reduction untuk waktu komputasi yang dibutuhkan adalah pada % feature selection yang semakin kecil integrasi peringkas dokumen otomatis sebagai feature reduction membutuhkan tambahan waktu komputasi tersendiri, akan tetapi pada proporsi feature selection yang semakin besar, % peringkas dokumen otomatis dapat menurunkan waktu komputasi yang digunakan.

DAFTAR PUSTAKA

[1] H. Al-mubaid and A.S. Umair, "A new text categorization technique using distributional clustering and learning logic," *IEEE Trans. Knowl. Data Eng*, vol. 18, 2006, pp. 1156-1165.

[2] R. Peter, S. G, D. G, & S. Kp, "Evaluation of SVD and NMF Methods for Latent Semantic Analysis," *InternationalJournal of Recent Trends in Engineering*, vol. 1, 2009, pp. 308-310.

[3] Rakesh Peter, Shivapratap G, Divya G & Soman KP, "Evaluation of SVD and NMF Methods for Latent Semantic Analysis," *International Journal of Recent Trends in Engineering ,Vol 1, No. 3, May 2009.*

[4] Ladda Suanmali, Naomie Salim & M Salem Binwahlan, " Automatic text summarization using feature based fuzzy extraction," *Jurnal teknologi Maklumat jilid 20. Bil 2, 2008.*

[5] Q. Yang, "Support vector machine for customized email filtering based on improving latent semantic indexing," *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, vol. 6, 2005, pp. 3787 - 3791.

[6] Hussam Dahwa Abdulla, Martin Polovincak & Vaclav Snasel, "Using a Matrix Decomposition for Clustering Data," *International Conference on Computational Aspects of Social Networks 2009 IEEE.*

[7] Wu, R., "Improved K-Modes Clustering Method Based on Chi-square Statistics," *International Conference on Granular Computing. doi: 10.1109/GrC.2010.66. IEEE 2010.*