KLASTERISASI, KLASIFIKASI DAN PERINGKASAN TEKS BERBAHASA INDONESIA

***Suwanto Raharjo*[1]**
***Edi Winarko*[2]**

[1]*Teknik Informatika, Fakultas Teknologi Industri, Institut Sains & Teknologi AKPRIND*
[2] *Ilmu Komputer, Fakultas Matematika danIlmu Pengetahuan Alam Universitas Gadjah Mada*
[1]*wa2n@akprind.ac.id,* [2]*edwin@ugm.ac.id*

Abstrak

Studi pustaka penelitian di bidang klasterisasi dan klasifikasi dokumen teks berbahasa Indonesia menunjukan bahwa penelitian bidang pemrosesan dokumen telah dimulai pada tahun 2000. Terdapat berbagai metode data mining untuk melakukan pengelompokan dokumen digunakan seperti single pass filtering, Naive Bayes, Hirarki dan metode lainnya. Penelitian ini akan melakukan survei paper penelitian data mining teks berbahasa Indonesia. Dari paper yang didapatkan terlihat bahwa sebagian besar topik penelitian data mining bertujuan adalah untuk melakukan pengelompokan suatu berita baik online maupun cetak berdasar atas acuan tertentu, penelitian lain ditujukanuntuk mengolah teks di media sosial seperti twitter. Artikel ini akan memperlihatkan metode yang digunakan dan tujuan dari paper dalam bidang klasterisasi,klasifikasi danperingakasan dokumen berbahasa Indonesia.

Kata Kunci: *klasterisasi, klasifikasi, peringkasan, bahasaIndonesia, data mining.*

PENDAHULUAN

Perkembangan dokumen berbasis teks khususnya melalui Internet menyebabkan jumlah dokumen menjadi sangat besar dan menyebabkan pencarian didalam dokumen berbasis teks menjadi sebuah pekerjaan yang tidak mudah. Perkembangan tersebut direspon dengan penelitian di bidang informatika khususnya di bidang pemrosesan dokumen teks berbahasa Indonesia. *Data mining* merupakan salah satu ilmu dalam bidang informatika yang mempelajari penambangan data, dokumenteks merupakan salah satu dokumen yang ditambang. *Data mining* sendiri diartikan sebagai ekstraksi atau penambangan pengetahuan dari suatu data dengan jumlah yang besar (Han, J., Kamber, M., dan Pei, J., 2006). Fungsi utama dari *data mining* adalah untuk menentukan suatu pola yang didapatkan dari penugasan *data mining* (Han, J., Kamber, M., dan Pei, J., 2006). Tujuan dari *data mining* beragam mulai dari pengklasifikasi, pengelompokan, pecarian, peringkasan dokumen dan lain sebagainya. Dokumen teks merupakan salah satu bentuk dokumen yang sering ditambang. Proses penambangan data sendiri bukan merupakan sebuah proses yang tunggal namun merupakan proses yang berkelanjutan, dimulai dari adanya data mentah yangdilakukan proses awal

diikuti dengan proses penambangan data dan menghasilkan keluaran yang diharapkan.

METODE PENELITIAN

Penelitian ini dilakukan dengan melakukan survei daripaper hasil penelitian yang terbit di jurnal nasional atau pun yang dipublikasikan dalam seminar nasional dengan topik bahasan bidang klasterisasi, klasifikasi dan peringkasan teks berbahasa Indonesia. Paper yang didapatkan kemudian dilihat metode-metode yang digunakan dalam proses penambangan data, baik *pre* proses dan *post* prosesnya. Survei juga melihat banyaknya data pengujian yang digunakan tingkat akurasi yang didapatkan. Dalam survei ini juga melihat secara umum metode penulisan paper yang dilakukan.

HASIL DAN PEMBAHASAN

Dalam bagian ini akan dilakukan pembahasan mengenai survei yang dilakukan dalam paper yang sudah ada. Hasil penelusuran pustaka mendapatkan bahwa penelitian di bidang pemrosesan dokumen teks berbahasa Indonesia dimulai pada tahun 2000 (Ridha, A., Adisantoso, J., dan Bukhari, F., 2000). Walaupun penelitian initidak secara khusus melakukan klasifikasi, klasterisasi atau peringkasan dokumen namun penelitian awal inisecara umum dapat disebutkan sebagai pengklasifikasian dokumen teks dengan membuat indeks. Hasil survei menemukan bahwa terdapat paper yang kurang baik dalam tata cara penulisannya seperti permasalahan penulisan perujukan daftar pustaka (sitasi) yang tidak sesuai dengan daftar pustaka yang disertakan.

A. Pemrosesan Awal

Dalam proses ini dilakukan persiapan dokumen untuk siap menjadi bahan penambangan data. Pada pemrosesan dokumen teks, bagian ini merupakan proses melakukan pembersihan, perapihan, pembakuan, koreksi, standarisasi atau bahkan modifikasi dari isi dokumen teks. Mempersiapkan data dari dokumen sumber dapat dilakukan dengan memperhatikan beberapa permasalahan berikut (Bramer, M., 2007): 1) formulasi standar; 2) tipe data; 3) pembersihan data; 4) nilai penganti; dan 5) penyesuian jumlah atribut. Pemrosesan awal ditujukan untuk membuat dokumen input lebih konsisten untuk dapat memfasilitasi representasi teks yang digunakan untuk dapat dianalisis dalam proses berikutnya (Aggarwal, C.C., dan Zhai, C., 2012). Beberapa metode yang digunakan untuk mempersiapkan data agar siap dilakukan proses selanjutnya seperti: *data cleaning, stemming, removing stop words* dan *Term Frequency Inverse Document Frequency (TF IDF)*.

B. Standarisasi dan Pembersihan data

Paper yang disurvei tidak menyebutkan secara eksplisit metode standarisasi dokumen yang digunakan, namun secara umum terdapat proses pemilihan sumber data. Sebagai contoh terdapat paper yang melakukan pemilihan berita harian online dari 3 surat kabar nasional pada bidang politik nasional dan internasional yang terbit dalam 3 hari kecuali hari minggu (Ridha, A., Adisantoso, J., dan Bukhari, F., 2000). Metode pemilihan sumber data lain adalah menggunakan satu sumber harian berita online (Februariyanti, H., dan Zuliarso, E., 2013) (Arifin, A.Z., dan Setiono, A. N., 2002) (Februariyanti, H., dan Zuliarso, E., 2012b) (Samodra, J., Sumpeno, S., dan Hariadi, M.,2009) (Februariyanti,

H., Zuliarso, E., dan Utomo, M.S., 2012b) (Aristoteles, 2013), file abstrak satu seminar ilmiah (Februariyanti, H. dan Winarko, E., 2010), satu jenis majalah (Darujati, C., dan Gumelar, A.B., 2012), satu jenis jurnal (Saputra, N., 2012), email (Nashir, H., Sunaryono, D., dan Munif, A., 2012), data *tweet* (Andhika, F.R., dan Widyantoro,D.H., 2012), atau menggunakan data dari peneliti lain (Permadi,Y., 2008) (Marlina, M., 2012).

Namun terdapat beberapa pula peneliti yang tidak menyebutkan secara spesifik asal sumber datanya (Heriyanto, 2011) (Februariyanti, H., 2013) (Mustaqhfiri, M., Abidin, Z., dan Kusumawati, R., 2012) (Zaman, B., dan Winarko, E., 2011) (Hamzah, A., 2009) (Marvin, C.W., dan Semuil, T., 2010) (Arifin, A.Z., Darwanto, R., Navastara, D.A., dan Ciptaningtyas, H.T., 2008) (Kurniawan, B., Effendi, S.,dan Sitompul, O.S., 2012) (Hamzah, A., Susanto, A., Soesianto, F.,dan Istiyanto, J.E., 2007) (Hamzah, A., 2012) (Februariyanti, H., dan Zuliarso, E., 2012a). Selain melakukan pemilihan sumber data, pengubahan dan pembersihan data sumber dilakukan pada sumber data. Beberapa mekanisme yang dilakukan untuk melakukan perubahan data seperti, mengubah data sumber menjadi huruf kecil (*lowercase*) (Ridha, A., Adisantoso, J., dan Bukhari, F., 2000) (Darujati, C., dan Gumelar, A.B., 2012) (Andhika, F.R.,dan Widyantoro, D. H., 2012) (Kurniawan, B., Effendi, S., dan Sitompul, O.S., 2012), membuang karakter yang tidak diperlukan misal tanda baca atau karakter lain (Ridha, A., Adisantoso, J., dan Bukhari, F., 2000) (Andhika, F.R., dan Widyantoro, D.H., 2012) (Arifin, A.Z., Darwanto, R., Navastara, D.A., dan Ciptaningtyas, H.T., 2008) (Samodra, J., Sumpeno, S., dan Hariadi, M.,2009) (Kurniawan, B., Effendi, S.,dan Sitompul, O.S., 2012), mengganti karakter tertentu dengan karakter lain (Andhika, F.R., dan Widyantoro,D.H., 2012) dan menghapus angka (Februariyanti, H., 2013) (Andhika, F.R., dan Widyantoro, D.H., 2012) (Arifin, A.Z., Darwanto, R., Navastara, D.A., dan Ciptaningtyas, H.T., 2008).

Demikian juga melakukan penghilangan kata-kata yang tidak berguna dalam klasifikasi yang sering disebut dengan *stopwords* atau *stoplist* (Februariyanti, H., 2013) (Februariyanti, H., dan Zuliarso, E., 2013) (Darujati, C., dan Gumelar, A.B., 2012) (Hamzah, A., 2009) (Andhika, F.R., dan Widyantoro, D.H., 2012) (Arifin, A.Z., Darwanto, R., Navastara, D.A., dan Ciptaningtyas, H.T., 2008) (Arifin, A.Z., dan Setiono, A.N., 2002) (Februariyanti, H., dan Zuliarso, E., 2012b) (Samodra, J., Sumpeno, S., dan Hariadi, M.,2009) (Februariyanti, H., Zuliarso, E., dan Utomo, M.S., 2012) (Saputra, N., 2012) (Nashir, H., Sunaryono, D.,dan Munif, A., 2012) (Permadi,Y., 2008) (Hamzah, A., Susanto, A., Soesianto, F., dan Istiyanto, J.E., 2007) (Hamzah, A., 2012). Sebagian besar paper tidak menyebutkan secara jelas apa saja *stopword* yang digunakan dalam penelitiannya hanya memberikan contoh beberapa kata sambung atau kata ganti. Sedangkan Paper (Samodra, J., Sumpeno, S., dan Hariadi, M.,2009) menuliskan menggunakan 758 kata *stopword*, paper (Februariyanti, H.,dan Zuliarso, E., 2012b) dan (Februariyanti, H.,dan Zuliarso, E., 2013) menggunakan 780 kata *stopword* yang merujuk pada sumber yang sama yakni (Tala, F., 2003). Memberikan pembatasan tegas pada data sumber misalkan disebutkan bahwa data merupakan dokumen teks ASCII (*American Standard Code for Information Interchange*) (Ridha, A., Adisantoso, J., dan Bukhari, F., 2000),

ASCII tanpa simbol-simbol dan tanpa rumus-rumus (Heriyanto, 2011).

C. Stemming

Stemming merupakan proses yang menyediakan pemetaan varian morfologi yang berbeda dari suatu katake akar katanya (*stem*) (Tala, F., 2003). Algoritma stemming adalah sebuah prosedur untuk mengurangi semua kata yang memiliki akar kata yang sama ke dalam bentuk yang umum (Lovins, J.B., 1968). Beberapa algoritma yang digunakan dikembangkan dari algoritma yang sudah ada misalkan dari algoritma *porter* (Ridha, A., Adisantoso, J., dan Bukhari, F., 2000) (Februariyanti, H., Zuliarso, E., dan Utomo, M.S.,2012) (Darujati, C., dan Gumelar, A.B., 2012) (Mustaqhfiri, M., Abidin, Z., dan Kusumawati, R., 2012) (Februariyanti, H., 2013), *confix-stripping stemmer* (Kurniawan, B., Effendi, S.,dan Sitompul, O.S., 2012), *purelyruled-based stemmer* (Februariyanti, H.,dan Zuliarso, E., 2013), tidak menjelaskan algoritma yang digunakan (Arifin, A.Z., dan Setiono, A.N., 2002) (Februariyanti, H. dan Winarko,E., 2010) (Darujati, C., dan Gumelar, A.B., 2012) (Saputra, N., 2012) (Nashir, H., Sunaryono, D., dan Munif, A., 2012) (Andhika, F.R., dan Widyantoro, D.H., 2012) (Arifin, A.Z., Darwanto, R., Navastara, D.A., dan Ciptaningtyas, H.T., 2008) (Februariyanti, H., dan Zuliarso, E., 2012a) (Zaman, B., dan Winarko, E., 2011) (Februariyanti, H., 2013), namun ada juga yang tanpa menggunakan proses stemming (Samodra, J., Sumpeno, S., dan Hariadi, M.,2009) (Permadi, Y., 2008) (Heriyanto, 2011). Beberapa paper tidak secara jelas menyebutkan apakah terdapat proses stemming di dalam penelitiannya (Saputra, N., 2012) (Marlina, M., 2012) (Hamzah, A., 2009) (Marvin, C.W.,dan Semuil, T., 2010) (Hamzah, A., 2012) (Aristoteles, 2013).

D. Perhitungan Bobot

Pada tahun 1972, Karen Sp¨arck Jones mempublikasikan dalam *Journal of Documentation* sebuah paper dengan judul *A statistical interpretation of term specificityand its application in retrieval*, sebuah pengukuran dari kekhususan sebuah *term* yang kemudian dikenal dengan nama *inverse document frequency* (IDF) (Jones, K.S.,1972). Pengukuran tersebut didasarkan pada perhitungan frekuensi kemunculan *term* dalam dokumen. Formula dasar dari pengukuran IDF adalah seperti tertampil dalam formula 1. Di mana terdapat N dokumen dalam sebuah koleksi dimana *term*t_i muncul sebanyak n_i kali (Robertson, S., 2004).

$$idf(t_i) = log\frac{N}{n_i} \qquad (1)$$

Kumpulan *term* yang sudah diekstrak akan direpresentasikan dalam bentuk *Vector Space Model* (VSM). Pembobotan dalam VSM tersebut menggunakan bobot *tf-idf* yang dirumuskan secara umum seperti dalam formula2.

$$\omega_{t,d} = tf_{td}.log\frac{N}{df_d} \qquad (2)$$

Perhitungan tersebut digunakan untuk menilai bobot hubungan *term* terhadap dokumen. Metode pembobotan dokumen dengan metode ini cukup banyak digunakan dalam penelitian yakni: (Ridha, A., Adisantoso, J., dan Bukhari, F., 2000) (Arifin, A.Z., dan Setiono, A.N., 2002) (Februariyanti, H., Zuliarso, E., dan Utomo, M.S.,2012) (Darujati, C., dan Gumelar, A.B., 2012) (Saputra, N., 2012) (Nashir, H., Sunaryono, D., dan Munif, A., 2012) (Hamzah, A., 2009) (Hamzah, A.,

Susanto, A., Soesianto, F.,dan Istiyanto, J.E., 2007) (Februariyanti, H., dan Zuliarso, E., 2012a) dan (Mustaqhfiri, M., Abidin, Z., dan Kusumawati, R., 2012).

Beberapa paper menggunakan pembobotan suatu kalimatdan tidak menilai bobot dokumen, pembobotan kalimatseperti ini digunakan dalam penelitian peringkasan suatu dokumen. Seperti penggunaan modifikasi *tf-idf* dengan menggunakan 3 kalimat sebagai wakil dokumen yang disebut dengan *term frequency index block frequency* (*tf-ibf*) sebagai metode untuk pembobotan kalimat, dimana*tf-ibf* secara umum sama dengan *tf-idf* dengan istilah dokumen digantikan dengan blok (Zaman, B.,dan Winarko, E., 2011). Pembobotan dalam peringkasan dokumen dapat pula menggunakan metode regresi logistik biner (Marlina, M., 2012) atau metode genetika (Aristoteles, 2013). Metode dengan menghitung jarak antara jumlah *n-gram*yang dihasilkan dilakukan dalam paper (Permadi, Y., 2008) untuk melihat kedekatan dokumen. Namun terdapat paper yang tidak menyebutkan secara detail perhitungan bobot antar dokumen yang digunakan (Februariyanti, H. dan Winarko, E., 2010) (Permadi,Y., 2008) dan (Hamzah, A.,2012) atau apakah terdapat pembobotan dokumen dalam penelitiannya (Samodra, J., Sumpeno, S., dan Hariadi, M., 2009) (Februariyanti, H., dan Zuliarso, E., 2012b) (Februariyanti, H.,dan Zuliarso, E., 2013) (Andhika, F.R.,dan Widyantoro, D.H., 2012) (Kurniawan, B., Effendi, S., dan Sitompul, O.S., 2012) (Arifin, A.Z., Darwanto, R., Navastara, D.A., dan Ciptaningtyas, H.T., 2008) (Marvin, C.W.,dan Semuil, T., 2010) (Heriyanto, 2011) dan (Februariyanti, H., 2013).

PEMROSESAN DATA MINING

Sumber data yang telah diolah akan memudahkan untuk proses data mining. Pengetahuan yang akan ditambang menentukan fungsi dari *data mining* yang akan dilakukan seperti (Han, J., Kamber, M., dan Pei, J., 2006): 1) karakterisasi; 2) diskriminasi; 3) asosiasi/korelasi; 4) klasifikasi/prediksi; dan 5) klastering.

A. Klasifikasi

Klasifikasi adalah proses menentukan suatu obyek kedalam suatu kelas atau kategori yang telah ditentukan. Penentuan obyek dapat menggunakan suatu model tertentu bebeapa model yang bisa digunakan antara lain: *classification (IF-THEN) rules*, *decision trees*, formulamatematika atau *neural networks* (Han, J., Kamber, M., dan Pei, J., 2006). Klasifikasi data atau dokumen dimulai dengan membangun aturan klasifikasi dengan algoritma klasifikasi tertentu menggunakan data training (tahapan ini sering disebut dengan tahapan pembelajaran) dan tahap pengujian algoritma dengan data testing.

1) Data Latih dan Uji

Data latih dan uji dari paper dan jumlah dokumen yang digunakan dalam paper cukup bervariasi seperti tertampil di tabel 1.

Tabel 1. Penggunaan Data Latih dan Uji

Data Pelatihan	Jumlah dokumen	Paper
10% - 90%	2.400, 4 kelas kategori	(Samodra, J., Sumpeno, S., dan Hariadi, M., 2009)
30 dan 60 dokumen latih dan 30 dokumen uji	90 ,6 kelas kategori	kategori(Marvin, C.W., dan Semuil, T., 2010)
90%	400,4 kelas klaifikasi	Kurniawan, B., Effendi, S., dan Sitompul, O.S., 2012
30% sampai dengan 90% dengan maks 100 data uji	1000 dokumen berita	(Hamzah, A., 2012)
405 data latih, 45 data uji	450 dokumen abstrak	
174 dan 75 dokumen uji	249, 3 kelas kategori	(Saputra, N., 2012)
1000 data latih dan 500 per kelas	7000, 7 kelas klasifikasi	(Andhika, F.R., dan Widyantoro, D.H., 2012)
89 data latih	374 dokumen	(Permadi,Y., 2008).

2) *Na¨ıve Bayes*

Na¨ıve Bayes Classifier (NBC) merupakan metode yang berdasarkan atas probabilitas bayes untuk melakukan pengelompokan data (Domingos, P., dan Pazzani, M.,1997). Paper (Samodra, J., Sumpeno, S., dan Hariadi, M.,2009) menyimpulkan bahwa metode ini cukup baik untuk melakukan klasifikasi bahasa Indonesia dengan dokumen training rendah (20%) dapat menghasilkan tingkat akurasi di atas 80% dan penghilangan *stopwords*tidak memberingkan pengaruh yang signifikan. Paper (Marvin, C.W., dan Semuil, T., 2010) menyebutkan bahwa terdapat tingkat kesalahan sebesar 16,67% dari 30 dokumen yang coba diklasifikasi menggunakan meoide ini. Sedangkan (Kurniawan, B., Effendi, S., dan Sitompul, O.S., 2012) tidak menyebutkan hasil yang didapat namun memberikan kesimpulan bahwa hasil semakin akurat jika data latih diperbanyak. Dalam (Hamzah, A., 2012) menyebutkan bahwa algoritma ini memiliki kinerja tinggi untuk klasifikasi dokumen teks, bahkan dalam dokumen berita sampai 91% sedangkan dokumen akademik sampai 82% dan penggunakan dokumen latih sebesar 50% dapat memberikan kinerja akurasi diatas 75%. Paper (Darujati, C., dan Gumelar, A.B., 2012) memperkuat pendapat dalam paper (Samodra, J., Sumpeno, S., dan Hariadi, M., 2009) dimana penghilangkan *stopwords* hanya memiliki pengaruh yang kecil dalam metode ini. Paper (Darujati, C., dan Gumelar, A.B., 2012) juga menyebutkan tingkat akurasi dapat mencapai lebih dari 87% dengan dokumen latih 100 dokumen. Paper (Andhika, F.R., dan Widyantoro, D.H., 2012) yang membandingkan 3 metode yakni naive bayes, *decisiontree* dan *Support Vector Machine* (SVM) memberikan hasil bahwa metode NBC tidak maksimal untuk dokumen pendek bahasa Indonesia dibandingkan dengan SVM.

3) Metode lain

Paper (Andhika, F.R.,dan Widyantoro, D.H., 2012) yang membandingkan 3 metode menunjukan bahwa SVM memliki kinerja yanglebih

baik dibandingkan dengan *Decision Tree* dan NBC untuk klasifikasi teks pendek. Beberapa metode lainyang digunakan dalam klasifikasi dokumen adalah metode *semantic smoothing* (Zhang, X., Zhou,X., dan Hu, X.,2006) dengan ekstraksi ciri menggunakan *chi-square* yang digunakan dalam paper (Saputra, N., 2012) menghasilkan hasil klasifikasi lebih baik daripada tidak menggunakan *chi-square* dan bekerja lebih baik pada dokumen yang panjang dengan tingkat akurasi sampai dengan 97.33%. Penggunaan perbandingan frekuensi *N-gram* suatu dokumen digunakan dalam paper (Permadi,Y., 2008) berkesimpulan bahwa penggunaan *tri-gram* menghasilkanakurasi tertinggi 81.25%. Jika dokumen sample yang digunakan pembelajaran terlalu sedikit maka akan berpengaruh pada tingkat akurasi, penggunaan ontology dapat digunakan sebagai metode klasifikasi dengan tanpa dokumen pembelajaran (Februariyanti, H.,dan Zuliarso, E., 2012a), namun dalam paper ini tidak menyimpulkan kinerjadari system.

B. Klastering

Klastering merupakan pembagian suatu data ke dalam suatu group yang memiliki kemiripan obyek. Beberapa metode yang dapat digunakan dalam klastering adalah (Bramer, M., 2007) :1) partisi; 2) hirarki; 3) density based; 4) grid based; 5) model based. Menghitung nilai kemiripan *term* merupakan proses yang dilakukan untuk memulai klastering data suatu dokumen. Similaritas, kemiripan atau jarak satu *term* dengan *term* yang lain atau dapat juga disebut kesamaan antar dokumen A dengan dokumen B dapat diukur dengan fungsi similaritas tertentu. Beberapa algoritma yang bisa digunakan untuk menghitung kemiripan teks adalah *Euclidean Distance*, *cosine similarity*, *Jaccard Coefficient*, *Pearson Correlation Coefficient* dan *Kullback-Leibler Divergence* (Huang, A., 2008).

Dalam survei terlihat bahwa *cosine similarity* merupakan metode paling banyak digunakan, pembahasan mengenai algoritma kemiripan ini dapat dibaca lebih lanjut di paper (Salton, G., dan Buckley, C., 1988). Perhitungan tingkat kemiripan dengan metode ini dilakukan oleh paper (Februariyanti, H., dan Zuliarso, E., 2013) (Mustaqhfiri, M., Abidin, Z., dan Kusumawati, R., 2012) (Zaman, B., dan Winarko, E., 2011) (Hamzah, A., 2009) (Arifin, A.Z., dan Setiono, A.N., 2002) (Februariyanti, H., dan Zuliarso, E., 2012b) (Februariyanti, H., Zuliarso, E., dan Utomo, M.S.,2012) (Februariyanti, H. dan Winarko, E., 2010) dan (Hamzah, A., Susanto, A., Soesianto, F., dan Istiyanto, J.E., 2007).

Metode lain yang dapat digunakan untuk mengukur similaritas adalah *Suffix Tree Similarity Measure* (Chim, H.,dan Deng, X., 2007) yang digunakan dalam paper (Arifin, A.Z., Darwanto, R., Navastara, D.A.,dan Ciptaningtyas, H.T., 2008). Metode *exact match*dan pemberian nilai dengan *flag similarity* digunakan dalam paper (Heriyanto, 2011). Namun terdapat juga paper yang tidak menyebutkan secara jelas algoritma yang digunakan (Nashir, H., Sunaryono, D., dan Munif, A., 2012).

1) *Single Pass*

Single Pass merupakan algoritma yang hanya melakukan proses pemeriksaan dokumen hanya sekali (Rieber, S.,dan Marathe, U.,1969), Paper (Arifin, A.Z., dan Setiono, A.N., 2002) dan (Februariyanti, H., Zuliarso, E., dan Utomo, M.S.,2012) menggunakan algoritma ini untuk mengelompokan dokumen berdasarkan suatu *event* (kejadian) tertentu dan menyimpulkan bahwa metode ini cukup handal. Paper lain adalah (Februariyanti, H.,dan Zuliarso, E., 2012b) dan (Februariyanti,

H.,dan Zuliarso, E., 2013) yang menggunakan algoritma ini untuk mengelompokan dokumen berdasar atas kesamaan topik. Sedangkan di dalam paper (Nashir, H., Sunaryono, D.,dan Munif, A., 2012) digunakan untuk mengelompokan isi email didapatkan tingkat akurasi sebesar 68.24%.

2) *Agglomerative Hierarchical Clustering*

Metode klastering dapat dikategorikan dari tipe struktur yangdihasilkan yakni metode non-hirarki dan hirarki hirarkhis(Februariyanti, H. dan Winarko,E., 2010), metode hirarki dapat berupa *agglomerative* dimana klaster dikumpulkan dari mulai satu klaster secara bertahap atau *divisive* yakni klaster dipecah dalam setiap tahap hirarkinya (Fraley, C.,dan Raftery, A.E.,1998). Dengan metode ini disebutkanmampu menghasilkan klaster berisi dokumen dengantopik yang sama menggunakan data dari dokumen suatu seminar nasional (Februariyanti, H. dan Winarko, E., 2010).

C. Peringkasan

Peringkasan merupakan salah satu metode dalam data mining yang digunakan untuk mendapatkan diskripsi kompak atas suatu bagian dari data (Fayyad, U., Piatetsky-Shapiro, G., dan Smyth, P.,1996). Paper (Marlina, M., 2012) menggunakan metode regresi logistik biner untuk melakukan peringkasan dokumen teks bahasa Indonesia dengan tingkat akurasi 42.84%. Metode *Maximum Marginal Relevance* digunakan dalam paper (Mustaqhfiri, M., Abidin, Z., dan Kusumawati, R., 2012) untuk melakukan peringkasan dokumen berbahasa Indonesia menghasilkan tingkat presisi 77%. Paper (Aristoteles, 2013) menggunakan algoritma genetika untuk melakukan peringkasan teks dengan 100 dokumen latih dan 50 dokumen uji, hasil peringkasan 30% didapatkan tingkat akurasi sebesar 47.46%.

SIMPULAN DAN SARAN

Dari hasil survei yang dilakukan dapat disimpulkan bahwa penelitian di bidang klasterisasi, klasifikasi dan peringkasan dokumen berbahasa Indonesia masih belum banyak dilakukan. Demikian juga metode yang digunakan dalam melakukan klasterisasi dan klasifikasi dokumen teks berbahasa Indonesia masih kurang beragam. Metode yang paling banyak digunakan dalam klasifikasi adalah *naive bayes* dan *single pass* di klasterisasi. Jumlah dokumen latih dan uji bervarisi di masing-masing paper, namun secara umum cukup banyak yang menggunakan dokumen latih dan uji dengan perbandingan 90% dan 10% dari total dokumen. Survei juga menunjukkan bahwa cukup banyak paper yang kurang memperhatikan masalah penulisan dan pengacuan daftar pustaka. Dari hasil survei tersebut maka diperlukan penelitian dalam bidang klasifikasi, klasterisasi dan peringkasan Bahasa Indonesia dengan metode yang lebih beragam.

DAFTAR PUSTAKA

Andhika, F.R., dan Widyantoro, D.H., "*Klasifikasi Topik Terhadap Teks Pendek pada Jejaring Sosial Twitter*", Jurnal Sarjana ITB bidang Teknik Elektro dan Informatika, Volume 1, Nomor. 3, 2012.

Aggarwal, C.C., dan Zhai, C., *Mining Text Data*. Springer, 2012.

Arifin, A.Z., dan Setiono, A.N., "*Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia*

dengan Algoritma Single Pass Clustering", Proceeding of Seminar on Intelligent Technology and Its Applications (SITIA), Teknik Elektro, 2002.

Arifin, A.Z., Darwanto, R., Navastara, D.A., dan Ciptaningtyas, H.T., "*Klasifikasi Online Dokumen Berita Dengan Menggunakan Algoritma Suffix Tree Clustering*", Prosiding Seminar Sistem Informasi Indonesia (SESINDO 2008). ITS, Surabaya, Volume 17, 2008.

Aristoteles, "*Penerapan Algoritma Genetika pada Peringkasan Teks Dokumen Bahasa Indonesia*", Prosiding Semirata 2013, vol. 1, no. 1, 2013.

Bramer, M., *Principles of Data Mining*. Springer, 2007.

Chim, H., dan Deng, X., "*A new suffix tree similarity measure for document clustering*" Proceedings of the 16th international conference on World Wide Web. ACM, 2007: 121–130.

Darujati, C., dan Gumelar, A.B., "*Pemanfaatan Teknik Supervised untuk Klasifikasi Teks Bahasa Indonesia*", Jurnal LINK, Volume 16, Nomor 1, Februari 2012: 5-1 s/d 5-8

Domingos, P., dan Pazzani, M., "*On the optimality of the simple Bayesian classifier under zero-one loss*", Machine learning, Volume 29, Nomor 2-3, 1997: 103–130.

Fayyad, U., Piatetsky-Shapiro, G., dan Smyth, P., "*From data mining to knowledge discovery in databases*", AI magazine, Volume 17, Nomor 3, 1996 : 37.

Februariyanti, H. dan Winarko, E., "*Klastering Dokumen Menggunakan Hierarchical Agglomerative Clustering*", Prosiding Seminar Nasional Sistem dan Teknologi Informasi (SNASTI) 2010, Volume 1, Nomor 1, 2010.

Februariyanti, H., dan Zuliarso, E., "*Klasifikasi Dokumen Berita Teks Bahasa Indonesia menggunakan Ontologi*", Jurnal Teknologi Informasi DINAMIK, Volume 17, Nomor 1, 2012a.

Februariyanti, H., dan Zuliarso, E., "*Algoritma Single Pass Clustering untuk Klastering Halaman Web*", Prosiding Seminar Nasional Komputer dan Elektro (SENOPUTRO) 2012, 2012b.

Februariyanti, H., Zuliarso, E., dan Utomo, M.S., "*Klastering Berita Online tentang Bencana dengan Algoritma Single Pass Clustering*", Prosiding Seminar Nasional MIPA UNNES, 2012.

Februariyanti, H., dan Zuliarso, E., "*Klastering Dokumen Berita dari Web menggunakan Algoritma Single Pass Clustering*", Jurnal Teknologi Informasi DINAMIK, Volume 18, Nomor 1, Januari 2013: 80-90.

Februariyanti, H., "*Perancangan Pengindeks Kata pada Dokumen Teks menggunakan Aplikasi Berbasis Web*", Jurnal Teknologi Informasi DINAMIK, Volume 18, Nomor 2, Juli 2013: 161-170.

Fraley, C., dan Raftery, A.E., "H*ow many clusters? Which clustering method? Answers via model-based cluster analysis*", The computer journal, Volume 41, Nomor. 8, 1998 : 578–588.

Hamzah, A., Susanto, A., Soesianto, F., dan Istiyanto, J.E., "*Clustering untuk Peningkatan Efektivitas Penyajian Informasi dari Mesin Pencari Teks*", Prosiding Seminar Nasional Teknologi 2007, 2007.

Hamzah, A., "*Temu Kembali Informasi Berbasis Kluster untuk Sistem Temu Kembali Informasi Teks Bahasa*

Indonesia", Jurnal Teknologi, Volume 2, Nomor 1, 2009.

Hamzah, A., "*Klasifikasi Teks dengan Naive Bayes Classifier (NBC)untuk Pengelopokan Teks Berita dan Abstract Akademis*", Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) 2012, IST AKPRIND, 2012.

Han, J., Kamber, M., dan Pei, J.,*Data Mining: Concepts andTechniques*. Morgan Kaufmann, 2006.

Heriyanto, "*Penggunaan Metode Exact Match untuk Menentukan Keeiripan Naskah Dokumen Teks*", Jurnal Telematika, Volume 8, Nomor 1, Juli 2011.

Huang, A., "*Similarity measures for text document clustering*", Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand, 2008 : 49–56.

Jones, K.S., "*A statistical interpretation of term specificity and its application in retrieval*", Journal of documentation, Volume 28, Nomor 1, 1972 : 11–21.

Kurniawan, B., Effendi, S., dan Sitompul, O.S., "*Klasifikasi Konten Berita Dengan Metode Text Mining*", Jurnal Online Dunia Teknologi Informasi, Volume 1, Nomor 1, 2012.

Lovins, J.B., "*Development of a Stemming Algorithm*", Mechanical Translation and Computational Linguistics, Volume 11, 1968.

Marlina, M., "*Sistem Peringkasan Dokumen Berita Bahasa Indonesia Menggunakan MetodeRegresi Logistik Biner*", SkripsiInstitut Pertanian Bogor, 2012.

Marvin, C.W., dan Semuil, T., "*Aplikasi Klasifikasi Dokumen Menggunakan Metoda Naive Baysian*", Prosiding Seminar Nasional Informatika 2010 (SemnasIF 2010), 2010.

Mustaqhfiri, M., Abidin, Z., dan Kusumawati, R., "*Peringkasan Teks Otomatis Berita Berbahasa Indonesia Menggunakan Metode Maximum Marginal Relevance*", Jurnal MATICS, 2012.

Nashir, H., Sunaryono, D., dan Munif, A., "*Perancangan dan Pembuatan Perangkat Lunak Pengelompokan Email secara Otomatis Memakai Klasifikasi Single Pass Clustering Berbasis Kerangka Kerja Play!*" Jurnal Teknik Pomits, Volume 1, Nomor 1, 2012 : 1–6.

Permadi, Y., "*Kategorisasi Teks Menggunakan N-gram untuk Dokumen Berbahasa Indonesia*", Skripsi, Institut Pertanian Bogor, 2008.

Ridha, A., Adisantoso, J., dan Bukhari, F., "*Pengindeksan Otomatis Densan Istilah Tunggal Untuk Dokumen Berbahasa Indonesia*", 2000.

Rieber, S., dan Marathe, U., "*The single pass clustering method*", Report ISR-16 to the National Science Foundation, Cornell University, Department of Computer Science, 1969.

Robertson, S., "*Understanding inverse document frequency: on theoretical arguments for IDF*", Journal of documentation, Volume 60, Nomor 5, 2004 :503–520.

Salton, G., dan Buckley, C., "*Term-weighting approaches in automatic text retrieval*", Information processing & management, Volume 24, Nomor 5, 1988 : 513–523.

Samodra, J., Sumpeno, S., dan Hariadi, M., "*Klasifikasi Dokumen Teks Berbahasa Indonesia dengan Menggunakan Na¨ıve Baye*s", Prosiding Seminar Nasional Electrical, Informatics, and It's

Educations. ITS Digital Repository, 2009.

Saputra, N., "*Klasifikasi Dokumen Bahasa Indonesia Menggunakan Semantic Smoothing dengan Ekstraksi Ciri Chi-square*", Skripsi, Institut Pertanian Bogor, Bogor, 2012.

Tala, F., "*A study of stemming effects on information retrieval in Bahasa Indonesia*", 2003.

Zaman, B.,dan Winarko, E., "*Analisis Fitur Kalimat untuk Peringkas Teks Otomatis pada Bahasa Indonesia*", Jurnal Indo CEISS, Volume 5, Nomor 2, 2011.

Zhang, X., Zhou, X., dan Hu, X., "*Semantic smoothing formodel-based document clustering*", Data Mining, 2006. ICDM'06. Sixth International Conference on. IEEE, 2006: 1193–1198.