KLASTERISASI, KLASIFIKASI DAN PERINGKASAN TEKS BERBAHASA INDONESIA

Suwanto Raharjo¹
Edi Winarko²

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

Abstrak


Kata Kunci: klasterisasi, klasifikasi, peringkasen, bahasa Indonesia, 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 direpons 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 ditambah. 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, peringkasakan dokumen dan lain sebagainya. Dokumen teks merupakan salah satu bentuk dokumen yang sering ditambah. Proses penambangan data sendiri bukan merupakan sebuah proses yang tunggal namun merupakan proses yang berkelanjutan, dimulai dari adanya data mentah yang dilakukan proses awal...
diikuti dengan proses penambangan data dan menghasilkan keluaran yang diharapkan.

**METODE PENELITIAN**

Penelitian ini dilakukan dengan melakukan survei dari paper hasil penelitian yang terbit di jurnal nasional atau pun yang dipublikasikan dalam seminar nasional dengan topik bahasan bidang klasifikasi, 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 pengujuan yang digunakan tingkat akurasi yang didapatkan. Dalam survei ini juga melihat secara umum metode penulisan paper yang dilakukan.

**HASIL DAN PEMBAHASAN**


**A. Pemrosesan Awal**


**B. Standarisasi dan Pembersihan data**


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

C. Stemming


D. Perhitungan Bobot


\[
idf(t_i) = \log \frac{N}{n_i} 
\]  

(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.

\[
\omega_{t,d} = tf_{td} \cdot \log \frac{N}{df_d} 
\]  

(2)


PEMROSESAN DATA MINING

Sumber data yang telah diolah akan memudahkan untuk proses data mining. Pengetahuan yang akan ditambah 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: \textit{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.
<table>
<thead>
<tr>
<th>Data Pelatihan</th>
<th>Jumlah dokumen</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>10% - 90%</td>
<td>2.400, 4 kelas kategori</td>
<td>(Samodra, J., Sumpeno, S., dan Hariadi, M., 2009)</td>
</tr>
<tr>
<td>30 dan 60 dokumen latih dan 30 dokumen uji</td>
<td>90,6 kelas kategori</td>
<td>kategori(Marvin, C.W., dan Semuil, T., 2010)</td>
</tr>
<tr>
<td>90%</td>
<td>400,4 kelas klasiifikasi</td>
<td>Kurniawan, B., Effendi, S., dan Sitompul, O.S., 2012</td>
</tr>
<tr>
<td>30% sampai dengan 90% dengan maks 100 data uji</td>
<td>1000 dokumen berita</td>
<td>(Hamzah, A., 2012)</td>
</tr>
<tr>
<td>405 data latih, 45 data uji</td>
<td>450 dokumen abstrak</td>
<td></td>
</tr>
<tr>
<td>174 dan 75 dokumen uji</td>
<td>249, 3 kelas kategori</td>
<td>(Saputra, N., 2012)</td>
</tr>
<tr>
<td>1000 data latih dan 500 per kelas</td>
<td>7000, 7 kelas klasiifikasi</td>
<td>(Andhika, F.R., dan Widyantoro, D.H., 2012)</td>
</tr>
<tr>
<td>89 data latih</td>
<td>374 dokumen</td>
<td>(Permadi, Y., 2008).</td>
</tr>
</tbody>
</table>

2) Naïve Bayes


3) Metode lain


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).


1) Single Pass

2) **Agglomerative Hierarchical Clustering**

Metode klasering dapat dikategorikan dari tipe struktur yang dihasilkan yakni metode non-hirarki dan hirarki hirarkis(Febrauriyanti, H. dan Winarko,E., 2010), metode hirarki dapat berupa *agglomerative* dimana klas terdastur dari mulai satu klas secara bertahap atau *divisive* yakni klas terpisah dalam setiap tahap hirarkinya (Fralely, C., dan Raftery, A.E.,1998). Dengan metode ini disebutkanmampu menghasilkan klas berisi dokumen dengantopik yang sama menggunakan data dari dokumen suatu seminar nasional (Febrauriyanti, H. dan Winarko, E., 2010).

C. Peringkasan


**SIMPULAN DAN SARAN**

Dari hasil survei yang dilakukan dapat disimpulkan bahwa penelitian di bidang klasering, klasifikasi dan peringkasan dokumen berbahasa Indonesia masih belum banyak dilakukan. Demikian juga metode yang digunakan dalam melakukan klasering dan klasifikasi dokumen teks berbahasa Indonesia masih kurang beragam. Metode yang paling banyak digunakan dalam klasifikasi adalah naive bayes dan single pass di klasering. Jumlah dokumen latih dan uji bervariasi 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, klasering dan peringkasan Bahasa Indonesia dengan metode yang lebih beragam.

**DAFTAR PUSTAKA**


Arifin, A.Z., dan Setiono, A.N., “Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia...
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
Suwanto dan Edi, Klasterisasi, Klasifikasi, dan … 399
Indonesia”, Jurnal Teknologi, Volume 2, Nomor 1, 2009.


Heriyanto, “Penggunaan Metode Exact Match untuk Menentukan Keeriran Naskah Dokumen Teks”, Jurnal Telematika, Volume 8, Nomor 1, Juli 2011.


Samodra, J., Sumpeno, S., dan Hariadi, M., “Klasifikasi Dokumen Teks Berbahasa Indonesia dengan Menggunakan Naive Bayes”, Prosiding Seminar Nasional Electrical, Informatics, and It’s


