PEMILIHAN FITUR DOKUMEN BAHASA INDONESIA UNTUK PENGELOMPOKAN DENGAN METODE K-MEANS

RAHMATIKA DEWI

DEPARTEMEN ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2013

PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi berjudul Pemilihan Fitur Dokumen Bahasa Indonesia untuk Pengelompokan dengan Metode K-Means adalah benar karya saya dengan arahan dari komisi pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor,Juli 2013

Rahmatika Dewi
NIM G64090082

ABSTRAK

RAHMATIKA DEWI. Pemilihan Fitur Dokumen Bahasa Indonesia untuk Pengelompokan dengan Metode K-Means. Dibimbing oleh JULIO ADISANTOSO

Temu kembali informasi memiliki dokumen yang sangat beragam dan berkembang secara pesat sehingga dibutuhkan adanya pengelompokan dokumen sehingga dengan banyaknya dokumen dapat memberikan informasi yang akurat, efisien dan efektif. Pengelompokan dokumen dapat dilakukan dengan teknik *clustering*. Teknik K-Means merupakan salah satu contoh dari *partitional clustering*. K-Means memiliki kesederhanaan dalam algoritme yang bertujuan untuk mendapatkan hasil pengelompokan yang sesuai. Pemilihan fitur *chi-square* dan IDF digunakan untuk mendapatkan kata unik sebagai penciri dari dokumen. Hasil pengelompokan dengan pemilihan fitur yang berbeda dibuat agar dapat dibandingkan untuk mendapatkan hasil yang diharapkan. Nilai akurasi yang didapatkan untuk pemilihan fiturIDF dan *chi-square* dengan ukuran 150 dokumen menggunakan *rand index* yaitu 26%, 75%.Nilai akurasi yang didapatkan untuk pemilihan fiturIDF dan *chi-square* dengan ukuran 457 dokumen menggunakan *rand index* yaitu 31%, 37%. Nilai akurasi yang didapatkan untuk pemilihan fitur *chi-square* dan IDF dengan ukuran 150 dokumen menggunakan *purity measure* yaitu 97%, 96%. Nilai akurasi yang didapatkan untuk pemilihan fitur IDF dan *chi-square* dengan 457 dokumen menggunakan *purity measure* yaitu 93%, 95%.

Kata kunci : K-Means, pengelompokan, pemilihan fitur.

ABSTRACT

RAHMATIKA DEWI. Indonesian Document Feature Selection to Grouping with K-Means. Supervised by JULIO ADISANTOSO

The field of document information retrieval has very diverse and rapidly-growing documents thereforethe need for methods to categorize documents effectively and efficiently increases. Categorizing documents can be performed using clustering techniques. This research uses the K-Means technique, one example of a partitioning clustering algorithm. K-Means is a simple algorithm that aims to get the appropriate grouping. Chi-square feature selection and the IDF were used to obtain the termsused as the unique identifiers of the documents. Clustering results with different feature selection techniques were made forcomparison to get the expected results.The accuracy values obtained for the IDF and the chi-square feature selection for data size 150 using rand index are26%, 75%, respectively.The accuracy values obtained for the IDF and the chi-square feature selection for data size 457 using rand index are31%, 37%, respectively. The accuracy values obtained for the IDF and the chi-square feature selection for data size 150 usingpurity measureare 97%, 96%, respectively. The accuracy values obtained for the IDF and the chi-square feature selection for data size 457 using rand index are 93%, 95%, respectively.

Keywords:K-Means, Clustering, Feature Selection

PEMILIHAN FITUR DOKUMENBAHASA INDONESIA UNTUK PENGELOMPOKAN DENGAN METODE K-MEANS

RAHMATIKA DEWI

Skripsi
sebagai salah satu syarat untuk memperoleh gelar
Sarjana Ilmu Komputer
pada
Departemen Ilmu Komputer

DEPARTEMEN ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2013

JudulPenelitian :Pemilihan Fitur Dokumen Bahasa Indonesia untuk Pengelompokan dengan Metode K-Means
Nama : Rahmatika Dewi
NIM : G64090082

Disetujui oleh

Ir. Julio Adisantoso M.Kom
Pembimbing

Diketahui oleh

Dr. Ir. Agus Buono, M.Si, M.Kom
Ketua Departemen

Tanggal Lulus:

PRAKATA

Puji dan syukur penulis panjatkan kehadirat Allah SWT yang senantiasa memberikan rahmat dan hidayah-Nya sehingga penulis dapatmenyelesaikan penelitian ini. Shalawat dan salam disampaikan kepada Nabi Muhammad SAW beserta keluarga, sahabat, dan pengikutnya yang tetap berada di jalan-Nya hingga akhir zaman.

Selama penelitian, penulis menyadari bahwa banyak pihak yang ikut membantu sehingga skripsi ini dapat diselesaikan, oleh karena itu penulis ingin menyampaikan ucapan terima kasih kepada:

1 Ayahanda Budi Mulya, Ibunda Siti Sapuroh Yulinda, Ibunda Anita Firda, Kakanda Fadly Nurmansyah atas doa, kasih sayang, dukungan, serta motivasi kepada penulis untuk penyelesaian penelitian ini.

2 Bapak Ir. Julio Adisantoso M.Kom selaku dosen pembimbing yang telah memberi banyak ide, saran, bantuan, serta dukungan sampai selesainya penelitian ini.

3 Bapak Sony Wijaya dan Bapak Ahmad Ridha selaku dosen penguji yang telah memberi masukan dan saran pada penelitian dan tugas akhir penulis.

4 Rekan-rekan satu bimbingan, Arini Daribti Putri, Fitria Rahmadina, Edo Apriyadi, Tedy Saputra, Ahmad Mansur Zuhdi dan Damayanti Elisabeth semoga lancar dalam melanjutkan penelitiannya.

5 Bagus Diponegoro, Sapariansyah, M.Haikal Dzulfikri, Galih Pribadi, Wisnu Febry Pradana, Srividola Wulandari, Aisyah Syahidah,Widya retno Utami, Listhia Dewi, Shitta Narendra, Rini Kurniawati dan rekan-rekan seperjuangan di Ilmu Komputer IPB angkatan 46 yang tidak dapat disebut satu persatu atas segala kebersamaan, bantuan, dukungan, serta kenangan bagi penulis selama menjalani masa studi. Teman-teman asrama Hesti,Bagas,Nola,Sari,Anggi dan Osis angkatan 19. Semoga kita bisa berjumpa kembali kelak sebagai orang-orang sukses.

6 Rekan – rekan guru dan staff NIC dan GEC yang selalu meringankan beban pikiran dalam menyelesaikan penelitian ini dengan keceriaan kalian.

Penulis berharap penelitian ini dapat memberikan manfaat, khususnya bagi peneliti Ilmu Komputer dan Institut Pertanian Bogor pada umumnya.

Bogor, Juli2013

Rahmatika Dewi

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

PENDAHULUAN

Latar Belakang

Temu kembali informasi merupakan bagian dari ilmu komputer yang berkaitan dengan pengambilan informasi dari dokumen-dokumen berdasarkan pada isi dan konteks dari masing masing dokumen. Temu kembali informasi adalah sebuah media layanan bagi pengguna untuk memperoleh informasi atau sumber informasi yang dibutuhkan oleh pengguna. Dalam memenuhi keinginan pengguna informasi yang diberikan harus akurat agar terpenuhi dengan baik. Keakuratan suatu data dapat kita lihat dalam nilai evaluasi yang tinggi. Dengan besarnya volume dokumen teks dibutuhkan sistem yang dapat mengekstrasi informasi sehingga waktu untuk mendapatkan informasi menjadi lebih efisien dan efektif.

Sistem pencarian dokumen membantu pengguna ketika ingin mengetahui informasi yang dicari secara terarah. Ketika pengguna ingin mengetahui kelompok dokumen yang memuat lokasi tertentu yang sama, dibutuhkan sistem pencarian yang memberikan informasi kepada pengguna yang ingin melakukan pengelompokan dokumen tertentu. Pengelompokan tersebut berdasarkan kemiripan tertentu dari sebuah dokumen yang dilakukan pada metode *clustering* dokumen. Salah satu cara untuk meningkatkan hasil temu kembali informasi adalah dengan menerapkan algoritme statistik, diantaranya adalah *clustering* dan *classification* (Dhillon dan Modha 2000). *Clustering* pada dokumen telah lama diterapkan pada sistem pencarian untuk efektifitas dari temu kembali informasi. *Clustering* pada umumnya digunakan dalam proses penemuan topik dari dokumen yang bertujuan untuk menghasilkan kelompok dokumen masing-masing.

Banyak terdapat metode *clustering* dengan pendekatan umum seperti *exclusive partitioning, agglomerative clustering, hierarchical clustering*.K-Means termasuk dalam pendekatan *exlusive partitioning*. Metode ini dipilih karena pengelompokan dokumen yang terdapat pada pencarian informasi sangat banyak dan belum terkelompok dengan baik. Teknik *clustering* K-Means digunakan karena kesederhanaannya dalam berbagai bidang untuk pengenalan pola dan analisis *cluster*. Algoritme pengelompokan K-Means adalah untuk membangun sebuah partisi dari beberapa *dataset* benda menjadi satu set *cluster* yang ditentukan. Setiap segmen dari *dataset*diwakili oleh pusat *cluster*. Maka dari itu penelitian ini bertujuan untuk menghasilkan pengelompokan pada dokumen dengan menggunakan *clustering* K-means. Permasalahan mendasar clustering dokumen adalah tingginya dimensi data. Beberapa metode untuk mengurangi dimensi ada dua cara untuk mengurangi dimensi data, yaitu *feature selection* dan *feature transformation*.

Pemilihan fitur merupakan suatu proses memilih subset dari setiap kata unik yang ada di dalam himpunan dokumen latih yang akan digunakan sebagai fitur di dalam klasifikasi dokumen (Manning *et al*. 2008). Keunikan suatu kata pada dokumen untuk pengelompokan menjadikan kata unik tersebut sebagai kata penciri dari dokumen. Kata penciri dapat diperoleh dari pemilihan fitur yang digunakan. Penggunaan pemilihan fitur yang banyak dipakai adalah *Document Frequency (*DF*)* dengan membuang batasan nilai (*threshold*) yang rendah dan

yang memiliki nilai tinggi akan digunakan. Pemilihan fitur DF sering digunakan dalam dimensi reduksi karena kata yang ada di dalam dokumen yang jarang muncul memberikan sedikit informasi yang spesifik pada dokumen dan tidakmempengaruhi kinerja secara keseluruhan. Untuk pengelompokan ini digunakan pemilihan fitur *chi-square* yang menghasilkan kata unik dari tiap dokumen sebagai penciri dengan memakai taraf nyata sebagai batasan nilai unik dari tiap kata dalam dokumen. Pemilihan fitur *chi-square* digunakan untuk meningkat nilai akurasi dari pengelompokan dokumen (Herawan 2011).

Perumusan Masalah

Adapun perumusan masalah pada penelitian ini adalah

1 Apakah pemilihan fitur *chi-square* mampu meningkatkan pengelompokan ?
2 Apakah metode *clustering* K-Means mampu mengelompokkan dokumen dengan baik?

Tujuan Penelitian

Tujuan penelitian ini adalah mengetahui peningkatan yang terjadi pada pengelompokan dokumen menggunakan pemilihan fitur *chi-square* dan kemampuan metode *clustering* K-Means dalam pengelompokan.

Manfaat Penelitian

Manfaat dari penelitian ini mengetahui kemampuan metode *clustering* K-Means dalam pengelompokan dokumen dan mengetahui kemampuan pemilihan fitur *chi-square* dalam pengelompokan dokumen.

Ruang Lingkup Penelitian

Ruang lingkup pada penelitian ini adalah dokumen yang digunakan adalah dokumen berbahasa Indonesia Laboratorium Temu kembali Informasi Departemen Ilmu Komputer IPB dan koleksi dokumen yang digunakan sebanyak 607 dokumen yang memiliki struktur XML (*Extensible Markup Language*).

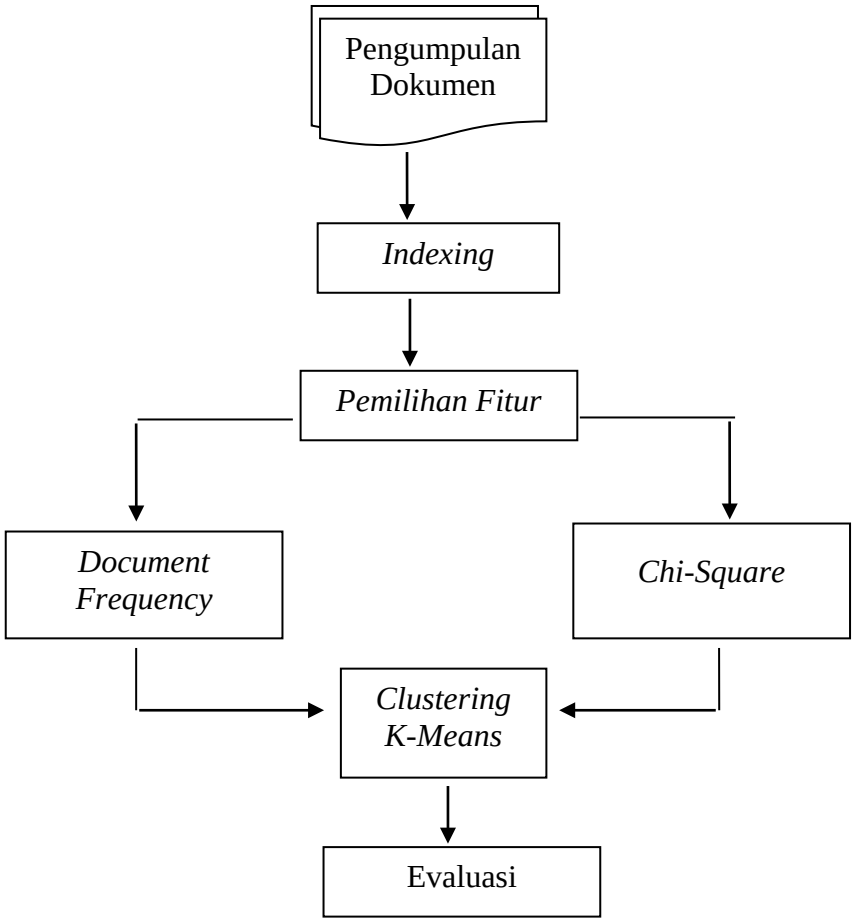
METODE

Tahapan dalam penelitian yang harus dilakukan yaitu pengumpulan dokumen, *indexing*, *clustering* K-Means, evaluasi.Metode penelitian dicantumkan pada Gambar 1.

Pengumpulan Dokumen

Penelitian ini menggunakan kumpulan dokumen yang berhubungan dengan pertanian. Kumpulan dokumen memiliki jumlah yang relatif sama dalam

tiap kelas. Dokumen yang digunakan dalam penelitian ini adalah milik laboratorium Temu Kembali Informasi IPB yang diambil dari sumber yang diantaranya surat kabar, jurnal pertanian dan internet. Dokumen yang digunakan dapat berupa format *plain teks* PDF, XML, HTML. Tetapi dalam sistem ini dokumen yang digunakan berupa XML(*Extensible Markup Language*). Kesalahan ejaan dan tata bahasa tidak diperbaiki karena merupakan isi dari dokumen dan tidak diubah. Format koleksi dokumen ada pada Gambar 2.



Gambar 1 Metode penelitian

```
<dok>
<id>2</id>
<judul>Budidaya Cabai</judul>
<deskripsi>Masalah utama budidaya cabai di lahan kering pegunungan dengan kemiringan adalah erosi tanah…</deskripsi>
</dok>
```

Gambar 2 Format dokumen

Indexing

Temu kembali berdasarkan konsep menunjukkan bahwa ide dalam dokumen lebih berhubungan pada konsep yang menggambarkan dokumen daripada kata-kata. Jadi, metode temu kembali harus mencocokkan konsep yang ditampilkan dalam *query* ke konsep yang ditampilkan dalam dokumen (Karypisdan Han 2000). *Indexing* adalah sebuah proses untuk melakukan ektraksi ciri yang terdapatpada kumpulan dokumen yang disediakan untuk dilakukan pencarian. Adapun tahapan dari pengindeksan meliputi tokenisasi, *stoplist*, *stemming*, dan pembobotan (Manning *et al*.2008).

Tokenisasi adalah proses pemenggalan (parsing) kata menjadi unit kecil yang disebut token dan pada saat yang sama membuang karakter tertentu seperti tanda baca yang terdapat dalam dokumen (Manning *et al*.2008). Token berupa masukan teks yang dibagi menjadi unit-unit kecil dapat berupa angka atau kata yang bertujuan untuk mempermudah dalam mengetahui frekuensi kemunculan tiap *token* pada suatu dokumen. Kata adalah sekumpulan karakter alfanumerik yang saling terhubung dan dipisahkan oleh *whitespace*, di antaranya adalah spasi, tab, dan *newline*.Dalam penelitian ini tanda baca dihilangkan dan mengubah kata menjadi *lowercase*.

Stopwords adalah daftar kata-kata yang dianggap tidak memiliki makna. Kata yang tidak ada di dalam *stopwords* dilanjutkan ke dalam proses selanjutnya, sedangkan kata yang ada di daftar *stopwords* dibuang. Pada umumnya kata yang masuk dalam *stopwords* adalah kata yang memiliki kemunculan yang sangat tinggi, yang sering muncul pada dokumen sehingga tidak dapat menjadi penciri dari dokumen.

Pembobotan katamencakup dua aspek yaitu lokal (*term frequency*) dan pembobotan global (*document frequency*). *Term frequency* (tf) adalah jumlah kemunculan setiap *term t* dalam sebuah dokumen *d* dan dinotasikan dengan $tf_{t,d.}$ sedangkan *document frequency* (df) adalah jumlah dokumen dalam koleksi suatu *term* . Untuk menghitung pembobotan suatu *term t* digunakan df yang dinotasikan df_t. Jika total seluruh dokumen dinotasikan dengan N maka ditetapkan *inverse document frequency* (Idf) dari sebuah *term* t yang disebut juga sebagai pembobotan global yaitu :

$$Idf_t = \log \frac{N}{df_t}$$

dengan df_t adalah jumlah dokumen yang mengandung *term t*. Nilai bobot dari suatu kata yang terpilih adalah perkalian antara kedua pembobotan yaitu :

$$tf_{t,d} \times Idf_t$$

dengan $tf_{t,d}$ adalah frekuensi *term t* pada dokumen d.

Pembobotan *term t* dalam dokumen *d* memiliki hubungan sebagai berikut :

1 Bobot tinggi ketika kemunculan t dalam jumlah dokumen yang kecil.
2 Lebih rendah ketika kemunculan *term* sedikit dalam sebuah dokumen atau muncul dalam banyak dokumen.

3 Paling rendah ketika muncul hampir diseluruh dokumen (Manning *et al*. 2008).

Pemilihan Fitur

Berdasarkan pernyataan dari Luhn (1958) atau yang biasa dikenal sebagai *Luhn Ideas*, bahwa kata-kata yang paling umum dan paling tidak umum adalah tidak signifikan untuk *indexing*. Kata-kata yang tidak dapat dijadikan sebagai penciri dari suatu dokumen adalah kata-kata yang kemunculannya sangat sering dan juga kata-kata yang kemunculannya sangat jarang pada sebuah dokumen sehingga kata-kata dengan frekuensi kemunculan yang cukup merupakan kata-kata yang paling baik digunakan sebagai penciri dari suatu dokumen. Pemilihan fitur merupakan proses menghilangkan beberapa fitur atau *term* yang kurang relevan untuk penentuan topik suatu dokumen. Pada seleksi fitur terdapat dua bagian yaitu *unsupervised* dan *supervised*. Keberadaan informasi awal pada kategori suatu dokumen yang menjadi berbeda antara *supervised* dan *unsupervised*. *Chi-square* adalah pemilihan fitur yang termasuk dalam bagian *supervised* yang mampu menghilangkan banyak fitur tanpa mengurangi tingkat akurasi sehingga dapat menghasilkan kata unik yang dapat menjadi penciri dari suatu dokumen. Penggunaan *chi-square* yang merupakan pemilihan fitur *supervised* dan *clustering* yang termasuk pengelompokan *unsupervised* bertujuan untuk mengetahui kemampuan *chi-square* untuk meningkatkan kemampuan pengelompokan dokumen. Pemilihan fitur *chi-square* berfungsi untuk menyeleksi *term* yang memiliki kontribusi dengan penentuan sebuah dokumen dan meningkatkan kinerja dari *clustering* dokumen. *Chi-square* dilakukan dengan cara membagi data menjadi dua yaitu data *training* sebesar 70% dan data *testing* sebesar 30%. Tabel kontingensi antara kata dengan kelas untuk perhitungan *chi-square*dapat dilihat pada Tabel 1.

Tabel 1Kontingensi kata dengan kelas

Predicted Class	*Actual Class*	
	Kelas= 1	Kelas= 0
Kata= 1	A	B
Kata= 0	C	D

Dari hasil perhitungan menggunakan Tabel 1 dimasukkan kedalam perhitungan untuk menghitung nilai *chi-square* pada suatu dokumen sebagai berikut :

$$x^2(t,c)= \frac{N(A*D-B*C)^2}{(A+B)*(C+D)*(A+C)*(B+D)}$$

dengan t merupakan kata yang sedang diujikan terhadap suatu kelas c, N merupakan jumlah dokumen latih, A merupakan banyaknya dokumen pada kelas c yang memuat kata t, B merupakan banyaknya dokumen yang tidak berada di c namun memuat kata t, C merupakan banyaknya dokumen yang berada di kelas c namun tidak memiliki kata t di dalamnya, serta D merupakan banyaknya dokumen yang bukan merupakan dokumen kelas c dan tidak memuat kata t.

Pemilihan fitur lain yang digunakan dalam penelitian ini adalah *Document Frequency* (DF). *Document Frequency* adalah jumlah dokumen yang mengandung suatu *term* tertentu. Tiap *term* akan dihitung nilai *Document Frequency*-nya (DF) lalu *term* tersebut diseleksi berdasarkan jumlah nilai DF. Jika nilai DF berada di bawah *threshold* yang telah ditentukan, maka *term* tersebut akan dibuang. *Term* pada DF yang lebih jarang muncul tidak memiliki pengaruh yang besar dalam proses pengelompokan dokumen. Pembuangan *term* yang jarang muncul pada tiap dokumen ini dapat mengurangi dimensi fitur yang besar pada sebuah dokumen.

Clustering K-Means

Clustering secara garis besar dibagi menjadi dua kelompok yaitu *hierarchical* dan *partitional*. *Hierarchical clustering* secara *rekursif*dapat menemukan *cluster* dengan cara *agglomerative* dan *divisive*. *Agglomerative* secara *rekursif* menggabungkan sepasang titik yang memiliki paling banyak kesamaan ke dalam satu *cluster* sehingga berbentuk hirarkikal. *Divisive* secara rekusrsif membagi titik dalam sebuah *cluster* menjadi *cluster* yang lebih kecil. *Partitional clustering* adalah algoritme menemukan semua *cluster* secara simultan sebagian bagian data dan tidak membentuk suatu hierarkikal (Jain 2009).

Berbeda dengan *association rule mining* dan *classification* dimana kelas data telah ditentukan sebelumnya, *clustering* melakukan pengelompokan data tanpa berdasarkan kelas data tertentu. Bahkan *clustering*dapat dipakai untuk memberikan label pada kelas data yang belum diketahui. Karena itu *clustering* sering digolongkan sebagai metode *unsupervised learning*. Prinsip dari *clustering* adalah memaksimalkan kesamaan antar anggota satu kelas dan meminimumkan kesamaan antar kelas/*cluster*.*Clustering* dapat dilakukan pada data yang memiliki beberapa atribut yang dipetakan sebagai ruang multidimensi. Banyak metode *clustering* yang digunakan untuk mengelompokkan dokumen ke dalam kelas yaitu K-Means, UPGMA, *Fuzzy* K-Means, *Bisecting* K-Means dan lain-lain. Metode yang digunakan dalam penelitian ini adalah K-Means karena algoritme *clustering* untuk mengenali pola dan menganalisis *cluster*. K-Means dapat dikatakan selalu menghasilkan *cluster* analisis sukses karena algoritme yang efisien. Berikut ini adalah algoritme K-Means untuk menemukan *K cluster* pada sebuah koleksi dokumen yaitu :

1 Menginisialisasikan *cluster* dengan *k-centroid*.
2 Masukan setiap dokumen ke dalam *cluster* yang paling cocok berdasarkan ukuran kedekatan dengan *centroid*.
3 Setelah semua dokumen masuk ke dalam *cluster*, maka pusat *centroid cluster* dihitung ulang berdasarkan dokumen yang ada di dalam *cluster* tersebut.
4 Jika *centroid* tidak berubah, maka proses selesai. Sebaliknya jika *centroid* berubah, maka hitung kembali ke proses 2.

Evaluasi

Evaluasi dalam pengelompokan bertujuan mencapai tingkat kesamaan intra-*cluster* (dokumen dalam *cluster* yang sama) dan rendah kesamaan antar *cluster* (dokumen dari *cluster* yang berbeda adalah berbeda) dengan nilai akurasi

yang baik. Evaluasi hasil *cluster* menggunakan *rand index* dan *purity measure*. Tabel *confusion matrix* untuk Rand Index dapat dilihat padaTabel 2.

Tabel 2 *Confusion matrix* untuk *rand index*

Predicted class	*Actual class*	
	Cluster yang sama	*Cluster* yang berbeda
Benar sama	A	C
Benar berbeda	B	D

dengan *A* adalah keputusan menempatkan dua dokumen yang mirip ke *cluster* yang sama, *B* adalah keputusan menempatkan dua dokumen yang tidak mirip ke *cluster* yang berbeda. *C adalah* keputusan menempatkan dua dokumen yang tidak mirip ke *cluster* yang sama. *D dalah* keputusan menempatkan dua dokumen yang mirip ke *cluster* yang berbeda.Akurasi dari pengelompokan *rand index* diperoleh dari formula:

$$\textit{Rand Index} = \frac{A + D}{A + B + C + D}$$

Hasil pengukuran (*performance metric*) dapat diperoleh dengan melihat hasil *rand index* (RI). Selain *rand index* pengukuran *cluster* juga dilakukan dengan menggunakan *purity measure* (PM). *Purity measure* adalah teknik evaluasi dari pengelompokan yang sederhana dan transparan. Untuk menghitung nilai dari *purity measure* adalah mengambil dokumen dari tiap *cluster* yang paling sering muncul kemudian keakuratan diukur dengan menghitung jumlah dokumen yang benar dan membaginya dengan seluruh jumlah dokumen. Akurasi pengelompokan *purity measure* diperoleh dari formula :

$$\textit{Purity Measure} = \frac{1}{N}\sum_{k}\max_{j}/\omega k \cap Cj/$$

dengan N sebagai jumlah seluruh dokumen. ωk sebagai set dari *cluster* jdan Cj adalah set dari kelas j. Dalam penelitian ini, pengelompokan dokumen yang telah dianggap benar adalah pengelompokan yang dilakukan dengan cara manual (Ramdani, 2011). Jadi evaluasi yang digunakan dengan perhitungan *Rand Index* dan *Purity Measure* dilakukan dengan cara dihitung secara manual.

Lingkungan Pengembangan Sistem

Penelitian ini menggunakan perangkat lunak dan perangkat keras dengan spesifikasi adalah sebagai berikut :

1 Perangkat Lunak :

- Sistem operasi Microsoft Windows 7 Ultimate 32-bit
- Notepad++ sebagai *code editor*
- Matlab R2008b dan *Library* K-Means

- Microsoft Office 2007 sebagai aplikasi yang digunakan untuk melakukan perhitungan dalam evaluasi sistem.

2 Perangkat Keras :

- Intel Pentium Core i5 @3.0 GHz
- *Memory*4096MB RAM
- *Harddisk* dengan kapasitas 320GB
- Monitor resolusi 1366 × 768 *pixel*
- *Mouse* dan *keyboard*

HASIL DAN PEMBAHASAN

Karakteristik Dokumen

Penelitian kali ini digunakan data dari Laboratorium Temu Kembali Informasi IPB yaitu dokumen ekofiologi dan agronomi, pemuliaan dan agronomi, proteksi (hama dan penyakit), tanaman obat dan hortikultura. Tema dari tiap kelompok dokumen tidak memiliki keterkaitan atau memiliki hubungan yang jauh dengan kelas lain.Dokumen yang digunakan berbahasa Indonesia. Seluruh dokumen berformat XML yang memili ekstensi *xml. Struktur tulisan yang terdiri atas dok, id, contentdapat dilihat pada Gambar 3.Seluruh dokumen dibagi menjadi 2 yaitu dengan pembagian seperti padaTabel 3.

```
<doc>
<docid>1</docid>
<content>Nama : Pandan Wangi Nama Latin </content>
</doc>
```

Gambar 3 Format dokumen XML

Tabel 3 Jumlah dokumen

Jumlah dokumen	Jumlah kelas	Nama kelas
457 (Herawan 2011) dan (Sari 2012)	2 kelas	1. Hortikultura 2. Tanaman Obat
150 (Ramadhina 2011)	3 kelas	1. Ekofiologi dan Agronomi 2. Pemuliaan dan Agronomi 3. Proteksi (Hama dan Penyakit

Jadi total keseluruhan dokumen menjadi 607 dokumen. Proses pada dokumen dilakukan pada teks yang berada di dalam struktur tulisan <doc> dan </doc> sehingga id dan kelas juga akan diproses tetapi memiliki perbedaan. Perbedaan untuk struktur tulisan id dan kelas ada pada proses penentuan dokumen

termasuk ke dalam kelas bagian mana. Untuk teks yang ada di dalam struktur tulisan <doc> dan </doc> akan dilakukan untuk proses untuk pembobotan nilai pada tiap kata di dalam dokumen.

Indexing

Pada tahap *indexing* dilakukan penghapusan *stopwords* dan tokenisasi untuk menghasilkan *term* yang sesuai. Seluruh kata di dalam dokumen dilakukan proses tokenisasi tetapi masih terdapat kata-kata yang termasuk ke dalam deret kata *stopwords*. Jumlah *term* awal memiliki jumlah yang lebih besar dibandingkan setelah dilakukan pengurangan *stopwords*. Total *term* setelah dilakukan pengurangan *stopwords* dan tokenisasi dapat dilihat pada Tabel 4.

Tabel 4 Hasil *stopwords* dan tokenisasi

Proses	Jumlah Kata	
Jumlah Dokumen	150	457
Tokenisasi	6802	12182
Stopwords	1584	7174

Matrik *document frequency* dan *chi-square* diperoleh dari hasil pencocokan kata yang terdapat pada hasil pemilihan fitur *document frequency* dan *chi-square* dengan kata pada hasil dari pembobotan Tf-Idf. Dengan demikian koleksi dokumen dapat dituliskan sebagai matrik kata-dokumen X adalah sebagai berikut :

$$X = \{X_{ij}\}\ i= 1,2,..t\ ;\ j =1,2,..\ N$$

dengan X_{ij} adalah bobot *term* i dalam dokumen ke j. *Document Frequency* menganggap setiap *term* memiliki tingkat kepentingan yang sama walaupun terdapat di berbagai dokumen. Hal ini berarti semakin banyak *term* tersebut terdapat di dalam dokumen yang berbeda, maka nilainya semakin besar dan memiliki pengaruh yang semakin besar pula pada *clustering* dokumen. Dengan menggunakan batasan nilai *chi-square* maka akan terjadi pemangkasan *term* pada suatu dokumen yang mempunyai nilai dibawah batas yang ditentukan. Perbandingan yang terdapat pada hasil *chi-square* dan DF terdapat pada jumlah *term* yang dihasilkan. Pada *chi-square* kata yang dihasilkan lebih sedikit daripada DF. *Chi-Square* menggunakan batasan pemilihan fitur sesuai hasil penelitian Saputra (2012) yaitu dengan nilai 6,63 atau dengan taraf $\alpha = 0{,}01$ akan menghasilkan nilai akurasi yang baik sehingga jumlah matrik yang akan diolah untuk *clustering* terdapat pada Tabel 5.

Tabel 5 Jumlah matrik dengan dimensi m × n

	Jumlah dokumen	
Pemilihan fitur	150	457
Document Frequency	45 × 766	137 × 6735
Chi-Square	45 × 199	137 × 1309

Clustering K-Means

Matrik yang dihasilkan dari proses sebelumnya dimasukkan ke dalam proses *clustering* pada Matlab. Matrik yang dimasukkan ke dalam fungsi merupakan matrik dua dimensi data. Proses *clustering* terdapat pada pemilihan fungsi K-means sesuai dengan algoritme yaitu penentuan *cluster*. Perhitungan jarak pada cluster dalam penelitian ini menggunakan perhitungan jarak yang sederhana, yaitu *euclidean distance*. *Euclidean distance* sering digunakan untuk menyatakan ketidaksamaan antara dua pola dengan menghitung jarak berdasarkan panjang vektor dari antar dokumen. Ukuran ini mengasumsikan bahwa antar sumbu koordinat dalam ruang vektor adalah saling bebas. Dalam vektor dokumen dimana koordinat adalah kata yang diekstrak dari koleksi dokumen dan dalam dokumen selalu ada kata yang kemunculannya tergantung pada kata yang lain. Fungsi yang dilakukan untuk menghasilkan IDX *cluster* ada pada Gambar 4.

```
[idx, ctrs] = kmeans(X, 2
                     'distance', 'sqEuclidean'
                     'start', 'cluster'
                     'options', 'opts')
```

Gambar 4 Contoh fungsi *clustering* K-Means

Hasil iterasi didapatkan dari perhitungan *centroid* sehingga *centroid* yang ditentukan diawal inisialisai diproses dan dilakukan beberapa kali iterasi sampai hasil *centroid* tidak berubah. Hasil iterasi pada *chi-square* jelas lebih sedikit dibandingkan dengan menggunakan *document frequency* karena jumlah matrik data yang dihasilkan juga lebih sedikit dibandingkan dengan *document frequency*. Hasil iterasi *clustering* dapat dilihat pada Tabel 6.

Tabel 6 Hasil iterasi *clustering*

	Jumlah dokumen	
Pemilihan fitur	150	457
Document Frequency	7	4
Chi-Square	4	2

IDX pada fungsi K-Means merupakan hasil dari pengelompokan dokumen. IDX yang digunakan pada *clustering* ini adalah dua dan tiga sesuai dengan pembagian tema pada dokumen. Hasil *clustering* menunjukkan bahwa pengelompokan yang menggunakan *chi-square* lebih baik dibandingkan dengan pemilihan fitur DF. Hasil *cluster chi-square* banyak masuk ke kelas yang seharusnya dibandingkan dengan DF. Kata unik yang ada di dalam dokumen yang dihasilkan oleh *chi-square* itulah yang menjadi peningkatan akurasi pengelompokan dokumen. Sedangkan pada proses DF hampir terdapat semua kata yang ada di dokumen setelah diproses *indexing*. Hasil IDX *clustering* dapat dilihat pada Lampiran 1.

Evaluasi

Perhitungan dengn tabel *confusion matrix* untuk evaluasi hasil dari pengelompokan dokumen menggunakan *rand index* dapat dilihat pada Lampiran 3.

Hasil tingkat akurasi dari hasil perhitungan evaluasi dalam bentuk persentasi dapat dilihat pada Tabel 7 dan Tabel 8.

Tabel 7 Hasil evaluasi *clustering* dengan *rand index*

Pemilihan fitur	150 dokumen	457 dokumen
Document Frequency	31 %	26 %
Chi-Square	37 %	75 %

Tabel 8 Hasil evaluasi *clustering* dengan *purity measure*

Pemilihan fitur	150 dokumen	457 dokumen
Document Frequency	93 %	97 %
Chi-Square	95 %	97 %

Hasil evaluasi lebih baik *purity measure* menghasilkan nilai akurasi yang tinggi dibandingkan dengan *rand index*. Hasil akurasi *purity measure* memiliki hasil evaluasi yang lebih tinggi karena pada perhitungan evaluasi purity measure diambil dokumen yang terbanyak yang ada di *cluster* tersebut, walaupun sudah ditentukan penentuan *cluster* secara acak tetapi pada penilaian evaluasi *purity measure* dokumen yang terbanyak dalam *cluster* tersebut maka dokumen tersebut memang masuk ke dalam *cluster* tersebut. Berbeda dengan perhitungan *rand index* yang memakai acuan penentuan *cluster* acak diawal sebelum diproses dan itu diperiksa kebenarannya. Penentuan *cluster* secara acak pada awal inilah yang dapat menyebabkan terjadinya kesalahan pada penentuan *cluster*. Jika penentuan awal *cluster* secara acak salah dan di evaluasi berbeda inilah yang menyebabkan nilai akurasi *rand index* menjadi kecil.

Dapat disimpulkan juga dari Tabel 7 bahwa pemilihan fitur *chi-square* mempunyai nilai yang lebih besar dibandingkan dengan pemilihan fitur DF karena pemilihan fitur *chi-square* membuat kata unik dari dokumen menjadi kata penciri sebuah dokumen sehingga kata yang didapatkan menjadi semakin sedikit dan akan menjadi penciri yang baik jika semakin sedikit kata unik pada tiap dokumen.

Clustering adalah *unsupervised* dan *document frequency* juga termasuk pemilihan fitur *unsupervised* sedangkan *chi-square* adalah pemilihan fitur *supervised*. Pengaruh *clustering* K-Means menggunakan pemilihan fitur sangat baik untuk meningkatkan akurasi walaupun berbeda jenis antara *unsupervised clustering* dan *supervised* pemilihan fitur tetapi baik jika digunakan dalam jumlah dokumen yang banyak agar dapat terlihat besar nilai akurasinya.

Dengan menggunakan data *training* dapat dilihat hasil evaluasi menggunakan *rand index* dan *purity measure* pada Tabel 9 dan Tabel 10. Tabel 9 dan Tabel 10 menunjukkan bahwa tingkat akurasi pada hasil data training mmenggunakan *chi-square* dan *document frequency* meningkat. Pengaruh peningkatan terjadi karena adanya *chi-square* yang meningkatkan hasil akurasi. Terbukti dengan menggunakan data *testing* dan data *training chi-square*dapat menghasilkan nilai akurasi dari pengelompokan suatu dokumen berbeda dengan hasil akurasi pada seluruh dokumen menggunakan *chi-square*. Hasil akurasi *chi-square* lebih kecil dibandingkan dengan menggunakan *document frequency*. Hasil akurasi dengan seluruh dokumen dapat dilihat pada Tabel 11 dan Tabel 12.

Tabel 9 Hasil evaluasi *clustering* data *training* dengan *rand index*

Pemilihan fitur	150 dokumen	457 dokumen
Document Frequency	35 %	30 %
Chi-Square	37 %	68 %

Tabel 10 Hasil evaluasi *clustering* data *training* dengan *purity measure*

Pemilihan fitur	150 dokumen	457 dokumen
Document Frequency	92 %	95 %
Chi-Square	92 %	98 %

Tabel 11 Hasil evaluasi *clustering* seluruh data dengan *rand index*

Pemilihan fitur	150 dokumen	457 dokumen
Document Frequency	33 %	69 %
Chi-Square	23 %	29 %

Tabel 12 Hasil evaluasi *clustering* seluruh data dengan *purity measure*

Pemilihan fitur	150 dokumen	457 dokumen
Document Frequency	97 %	98 %
Chi-Square	89 %	92 %

SIMPULAN DAN SARAN

Dengan hasil *clustering* pada kedua bagian dokumen dapat dilhat bahwa pengaruh yang terjadi pada pengelompokan dokumen dapat berubah. Pengaruh terjadi pengelompokandapat terjadi pada pengaruh kesesuaian kelas, keragaman dokumen dan pemilihan fitur yang dipilih.

Saran dalam penelitian ini menggunakan dokumen dengan jumlah kelas yang lebih banyak dan menggunakan metode *clustering* yang lain agar dapat dibandingkan secara lebih detail. Pemilihan fitur yang lain juga dapat menghasilkan perbedaan hasil akurasi pada suatu dokumen. Jadi dengan menggunakan metode pemilihan fitur lain dapat dilihat dan dibandingkan hasilnya.

DAFTAR PUSTAKA

Dhillon S I, Modha D S. 2000. *Concept Decompositions for Large Sparse Text Data using Clustering*.Kluwer Academic Publishers.

Herawan Y. 2011. Ekstraksi Ciri Dokumen Tumbuhan Obat Menggunakan *Chi-Kuadrat*denganKlasifikasi *Naive Bayes*[skripsi]. Bogor (ID). Departemen Ilmu Komputer. Insitut Pertanian Bogor.

Jain A K. 2009.*Data Clustering*: *50 Years Beyond K-Means*.*Department of Computer Science and Engineering*.Michigan State University. Michigan.

Karypis G, Han E. 2000. Concept Indexing: A Fast Dimensionally Reduction Algorithm with Apllications to Document Retrieval & Categorization. Computer Science and Engineering.University of Minnesota. Minneapolis.

Luhn HP. 1958. The automatic of literature abstracts. *IBM Journal of Research and Development*. 2(2): 159-165.

Manning CD, Raghavan P, Schütze H. 2008. *An Introduction to Information Retrieval*. Cambridge(UK): Cambridge University Press.

Ramadhina A. 2011. Klasifikasi Dokumen Bahasa Indonesia menggunakan Metode *Semantic Smoothing*. Bogor (ID). Departemen Ilmu Komputer. InstitutPertanian Bogor.

Ramdani H. 2011. *Clustering* Konsep Dokumen Berbahasa Indonesia menggunakan Bisecting K-Means [skripsi]. Bogor (ID). Departemen Ilmu Komputer. Institut Pertanian Bogor.

Saputra N. 2012.Klasifikasi Dokumen Bahasa Indonesia mengunakan *Semantic Smoothing* dengan Ekstraksi Ciri *Chi-Square* [skripsi]. Bogor (ID). Departemen Ilmu Komputer. Insitut Pertanian Bogor.

Sari PD. 2012.Metode pembobotan kata berbasis sebaran untuk temu kembali informasi dokumen bahasa Indonesia[skripsi]. Bogor (ID) : Institut Pertanian Bogor.

LAMPIRAN

Lampiran 1 Hasil IDX *clustering* 457 dokumen

No. Dok	Nomor *Cluster*	457 Dokumen	
		Document Frequency	*Chi-Square*
1	1	1	2
2	1	1	2
3	1	1	2
4	1	1	2
5	1	1	2
6	1	2	1
7	1	2	1
8	1	1	2
9	1	1	2
10	1	1	2
11	1	1	2
12	1	1	2
13	1	1	2
14	1	1	2
15	1	2	1
16	1	2	1
17	1	1	2
18	1	1	2
19	1	1	2
20	1	1	2
21	1	1	2
22	1	1	2
23	1	1	2
24	1	1	2
25	1	1	2
26	1	1	2
27	1	1	2
28	1	1	2
29	1	1	2
30	1	1	2
31	1	1	2
32	1	1	2
33	1	1	2
34	1	1	2
35	1	1	2
36	1	1	2
37	1	1	2
38	1	1	2
39	1	1	2
40	2	1	2

No. Dok	Nomor *Cluster*	457 Dokumen	
		Document Frequency	*Chi-Square*
41	2	1	2
42	2	1	2
43	2	1	2
44	2	1	2
45	2	1	2
46	2	1	2
47	2	1	2
48	2	1	2
49	2	1	2
50	2	1	2
51	2	1	2
52	2	1	2
53	2	1	2
54	2	1	2
55	2	1	2
56	2	1	2
57	2	1	2
58	2	1	2
59	2	1	2
60	2	1	2
61	2	1	2
62	2	1	2
63	2	1	2
64	2	1	2
65	2	1	2
66	2	1	2
67	2	1	2
68	2	1	2
69	2	1	2
70	2	1	2
71	2	1	2
72	2	1	2
73	2	1	2
74	2	1	2
75	2	1	2
76	2	1	2
77	2	1	2
78	2	1	2
79	2	1	2
80	2	1	2
81	2	1	2
82	2	1	2
83	2	1	2
84	2	1	2
85	2	1	2

No. Dok	Nomor *Cluster*	457 Dokumen	
		Document Frequency	*Chi-Square*
86	2	1	2
87	2	1	2
88	2	1	2
89	2	1	2
90	2	1	2
91	2	1	2
92	2	1	2
93	2	1	2
94	2	1	2
95	2	1	2
96	2	1	2
97	2	1	2
98	2	1	2
99	2	1	2
100	2	1	2
101	2	1	2
102	2	1	2
103	2	1	2
104	2	1	2
105	2	1	2
106	2	1	2
107	2	1	2
108	2	1	2
109	2	1	2
110	2	1	2
111	2	1	2
112	2	1	2
113	2	1	2
114	2	1	2
115	2	1	2
116	2	1	2
117	2	1	2
118	2	1	2
119	2	1	2
120	2	1	2
121	2	1	2
122	2	1	2
123	2	1	2
124	2	1	2
125	2	1	2
126	2	1	2
127	2	1	2
128	2	1	2
129	2	1	2
130	2	1	2
131	2	1	2

No. Dok	Nomor *Cluster*	457 Dokumen	
		Document Frequency	*Chi-Square*
132	2	1	2
133	2	1	2
134	2	1	2
135	2	1	2
136	2	1	2
137	2	1	2

Lampiran 2 Hasil IDX *clustering* 150 dokumen

No. Dok	Nomor *Cluster*	150 dokumen	
		Document Frequency	*Chi-Square*
1	1	3	3
2	1	2	1
3	1	3	3
4	1	3	3
5	1	3	3
6	1	3	3
7	1	3	3
8	1	3	3
9	2	3	3
10	2	3	2
11	2	3	3
12	2	3	3
13	2	3	3
14	2	3	3
15	2	3	3
16	2	1	3
17	3	1	3
18	3	3	3
19	3	3	3
20	3	3	3
21	3	3	3
22	3	3	3
23	3	3	3
24	3	3	3
25	1	3	3
26	1	3	3
27	1	3	3

No. Dok	Nomor Cluster	150 dokumen	
		Document Frequency	*Chi-Square*
28	1	3	3
29	1	3	3
30	1	3	3
31	1	3	3
32	2	3	3
33	2	3	3
34	2	3	3
35	2	3	3
36	2	3	3
37	2	3	3
38	2	3	3
39	3	3	3
40	3	3	3
41	3	3	3
42	3	3	3
43	3	3	3
44	3	3	3
45	3	3	3

Lampiran 3 *Confusion matrix* untuk perhitungan evaluasi *rand index*

Perhitungan untuk *document frequency* 457 dokumen

Predicted class	*Actual class*	
	Cluster yang sama	*Cluster* yang berbeda
Benar sama	35	4
Benar berbeda	98	0

Perhitungan untuk *chi-square* 457 dokumen

Predicted class	*Actual class*	
	Cluster yang sama	*Cluster* yang berbeda
Benar sama	4	34
Benar berbeda	0	98

Perhitungan untuk *document frequency* 150 dokumen

Predicted class	*Actual class*	
	Cluster yang sama	*Cluster* yang berbeda
Benar sama	0	15
Benar berbeda	16	14

Perhitungan untuk *chi-square* 150 dokumen

Predicted Class	*Actual class*	
	Cluster yang sama	*Cluster* yang berbeda
Benar sama	1	14
Benar berbeda	14	16

RIWAYAT HIDUP

Penulis lahir di kota Jakarta 21 tahun lalu pada tanggal 29Oktober 1991 sebagai anak kedua dari pasangan Budi Mulya dan Siti Sapuroh Yulinda. Penulis sekolah pendidikan dasar sampai menengah atas di Kota Jakarta.Penulis merupakan lulusan SMA Negeri 98 Jakarta (2006-2009), SMP Negeri 217 Jakarta (2003-2006), dan SDN 08 Baru Cijantung Jakarta (1997-2003).

Saat ini penulis sedang menyelesaikan studi S1 di Departemen Ilmu Komputer, Fakultas MIPA, Institut Pertanian Bogor sejak tahun 2009.Penulis sekarang menjadi guru honorer di salah satu bimbingan belajar di Kota Bogor.Selain itu, penulis melaksanakan kegiatan Praktik Kerja Lapangan di Bank Indonesia pada tahun 2012.