EFEKTIVITAS PENGGUNAAN *STOPLIST* KATA UMUM DARI DOKUMEN HASIL KLASIFIKASI *PRETOPOLOGY*

Mohammad Mastur[*], Fika Hastarita Rachman[**], Firdaus Solihin[***]
Program Studi Teknik Informatika, Fakultas Teknik, Universitas Trunojoyo
Jl. Raya Telang PO. BOX 2, Kamal, Bangkalan, Madura, 69162
E-Mail: [*]dmas89@gmail.com, [**]hastarita.fika@gmail.com,
[***]fsolihin@gmail.com

ABSTRAK

Dokumen teks bahasa Indonesia sangat melimpah dan setiap waktu bertambah. Dalam proses pencarian, banyak dokumen yang dihasilkan menjadi kurang relevan jika tidak sesuai dengan keinginan pengguna. *Stoplist* merupakan kumpulan kata yang "tidak relevan", namun sering muncul dalam dokumen. Kata umum juga sering digunakan pada bidang tertentu sehingga dimungkinkan untuk dokumen sebidang akan ada kata umum yang sering muncul. Pada sistem temu kembali informasi, asumsi yang ada adalah dengan menghapus *stoplist*, maka mengurangi jumlah kata yang akan diproses. Adanya kata umum bidang, memungkinkan adanya mengurangan jumlah kata yang akan diproses juga. Dalam penelitian ini dilakukan ekstraksi kata umum dari dokumen hasil klasifikasi dan melakukan perbandingan efektifitas antara aplikasi pencarian-1 menggunakan penghapusan *stoplist* dengan aplikasi pencarian-2 menggunakan penghapusan *stoplist* dan kata umum. Hal ini dilakukan untuk mengetahui model pencarian dengan tingkat relevansi dan waktu proses pencarian dokumen yang lebih tinggi. Hasil uji coba klasifikasi *pretopology* dengan 25 dokumen teknik, 25 ekonomi dan 25 pertanian diperoleh nilai rata-rata *recall* dan *precision* sebesar 90% dan 76%. Dan uji coba pencarian dengan 6 *query* terhadap 746 dokumen pada aplikasi pencarian-1 diperoleh nilai rata-rata *f-measure* dan waktu proses adalah 30.6% dan 0.239 detik. Sedangkan aplikasi pencarian-2 dengan *threshold* kata umum 1% adalah 76.5% dan 0.098 detik. Sehingga dapat dikatakan bahwa aplikasi pencarian-2 (dengan menggunakan penghapusan *stoplist* dan kata umum) lebih efektif dari pada aplikasi pencarian-1.

Kata kunci: Sistem temu kembali informasi, *Stoplist*, Klasifikasi *Pretopology*, Kata Umum.

1. Pendahuluan

Jumlah dokumen dalam bentuk teks mengalami peningkatan yang cukup signifikan dalam kurun waktu terakhir ini. Hal ini membuat proses pencarian dokumen dengan beribu data yang diperoleh kurang relevan atau tidak memiliki nilai. Sehingga sangat penting untuk mengorganisir dan mengklasifikasi dokumen secara otomatis. Ada dua varian dalam pengklasifikasian dokumen: *Clustering* dan pengklasifikasian dokumen. *Clustering* dokumen yaitu proses pengumpulan dokumen serupa ke dalam kelompok, di mana kesamaan adalah beberapa fungsi pada dokumen [1]. Sedangkan pengklasifikasian dokumen bertujuan membandingkan dan memasukkan dokumen baru ke dalam kelompok kategori berdasarkan struktur

kelompok yang sudah diketahui sebelumnya.

Sistem temu kembali informasi merupakan cara atau proses penyajian dokumen agar sesuai dengan apa yang dicari pengguna [2]. Dengan adanya pengklasifikasian dokumen maka akan membantu pengguna dalam pencarian dokumen dengan ditmpilkannya kategori terkait saja. Pada proses temu kembali informasi proses penghapusan *stopword* pada *preprocessing* menggunakan *stoplist* Indonesia yang bersifat umum. Sedangkan kata umum juga sering muncul ada bidang tertentu. Oleh kerena itu perlu adanya penelitian untuk melihat perbedaan hasil pengklasifikasian dokumen dengan penghapusan *stoplist* terhadap hasil pengklasifikasian dengan penghapusan *stoplist* dan kata umum. Kata umum diperoleh dari masing-masing bidang kemudian digunakan pada *preprocessing*. Diharapkan dengan penghapusan *stoplist* dan kata umum, akan mempercepat waktu proses pencarian dokumen dan meningkatkan relevansi hasil pencarian dokumen.

2. Tinjauan Pustaka

Pada tinjuan pustaka ini akan dibahas mengenai teori yang akan digunakan dan pemaparan mengenai penelitian terdahulu.

2.1 *Stoplist*

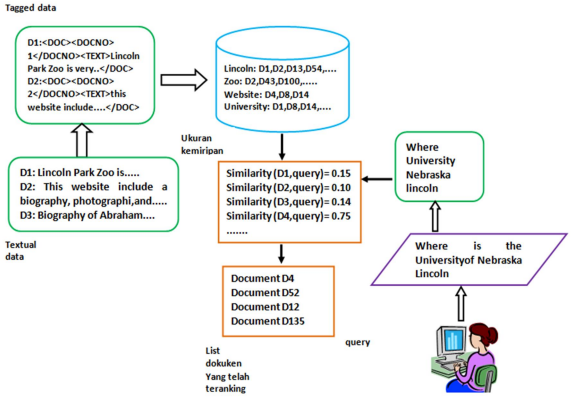
Kata fungsi dan kata sambung pada berbagai bahasa biasanya tidak memiliki arti bagi seseorang yang melakukan pencarian informasi berbasis kata kunci namun sering muncul pada banyak kalimat. Kata-kata seperti di, yang, dan pada bahasa Indonesia disebut sebagai *stopwords*.

Pada sebuah sistem temu kembali informasi, proses penghilangan *stopwords* (*stopwords removal*) perlu dilakukan. Sekumpulan *stopwords* sering disebut sebagai *stoplist* berisi sekumpulan kata yang tidak relevan yang sering muncul dalam dokumen. Proses penghilangan *stopwords* dilakukan pada hasil parsing sebuah dokumen teks dengan cara membandingkan dengan *stoplist* yang ada.

Pada penelitian ini, *stoplist* yang digunakan adalah *stoplist* untuk bahasa Indonesia hasil penelitian Tala [3].

2.2 Sistem Temu Kembali Informasi

Sistem temu kembali informasi bertujuan menemukan dokumen yang relevan terhadap permintaan pengguna dengan mencocokkan kata kunci yang dimasukkan oleh pengguna dengan dokumen yang ada. Arsitektur dari sebuah sistem temu kembali informasi diperlihatkan pada Gambar 1.



Gambar 1. Arsitektur Umum Sistem Temu Kembali Informasi

Inti dari Gambar 1 adalah sebagai berikut [4]:

1. *Database*, didefinisikan sebagai kumpulan dokumen teks.
2. Dokumen, terdiri dari urutan istilah bahasa alami yang mengekspresikan ide-ide tentang topik tertentu.
3. *Term*, didefinisikan sebagai unit semantik, frase, atau kata (atau

lebih tepatnya kata dasar).

4. *Query*, yang merupakan suatu permintaan untuk dokumen yang menyangkut topik tertentu yang menarik bagi pengguna.

2.2.1. Preprocessing

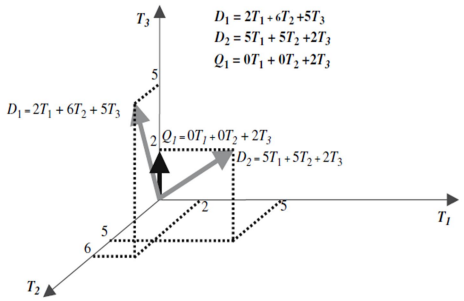
Pada tahap preprocessing ini dilakukan pembuatan Index dari sekumpulan dokumen yang diproses. Index merupakan himpunan istilah (term) yang ada dalam dokumen untuk menunjukkan isi atau topik dari dokumen.

Terdapat lima langkah pembangunan *inverted index* [5], yaitu:

1. Pengumpulan dokumen yang akan diindeks dan penghapusan format dan *markup* dari dalam dokumen.
2. Tokenisasi atau memisahkan rangkaian istilah dalam sebuah kalimat, paragraf menjadi potongan kata tunggal. Selain memisahkan kata atau istilah, pada tahap ini juga dilakukan pembersihan karakter yang tidak berarti seperti tanda baca serta mengubah karakter dalam token yang dihasilkan menjadi huruf kecil.
3. Melakukan praproses lingustik dari token-token yang dihasilkan atau disebut dengan istilah penyaringan (*filtration*). Pada langkah inilah dilakukan proses penghapusan *stopword* dengan menentukan daftar *stoplist* atau disebut dengan pustaka *stopwords*. Salah satu cara menyusun *stoplist* adalah dengan merangking frekuensi kemunculan term dari yang paling sering muncul dan memasukkan istilah yang sering muncul dalam *stopwords*.
4. Proses *stemming* atau mengembalikan bentuk asal dari setiap istilah yang diperoleh dengan menghilangkan awalan atau akhiran. Pada penelitian ini menggunakan teknik *Enhanced Confix Stripping* [6].
5. Pengindeksan istilah dengan memberikan bobot pada setiap istilah yang digunakan. Teknik pemberian bobot ini dapat menggunakan teknik pembobotan lokal, global atau kombinasi. Salah satu teknik kombinasi dikenal sebagai teknik TF-IDF merupakan kombinasi dari perkalian bobot lokal *term frequency* dan global *inverse document frequency*.

2.2.2. *Vector Space Model*

Vector Space Model (VSM) adalah teknik yang digunakan untuk mewakili dokumen dan permintaan sebagai vektor dalam ruang multidimensi, yang dimensinya adalah istilah yang digunakan untuk membangun indeks untuk mewakili dokumen [7]–[9]. Ini adalah teknik yang paling banyak digunakan untuk pengambilan informasi karena kesederhanaannya; efisiensi atas koleksi dokumen besar dan sangat menarik untuk digunakan. Efektivitas VSM sebagian besar tergantung pada bobot istilah yang diterapkan pada jangka waktu vektor dokumen. VSM memiliki tiga fase: (1) fase pengindeksan dokumen, di mana istilah-istilah yang mengandung konten diekstraksi dari teks dokumen; (2) pembobotan persyaratan yang diindeks untuk meningkatkan pengambilan dokumen yang relevan bagi pengguna; dan (3) membuat peringkat dokumen sehubungan dengan permintaan berdasarkan ukuran kesamaan. Gambar 2 menunjukkan contoh khas Model Ruang Vektor untuk dua dokumen, tiga istilah dan permintaan.



Gambar 2. Contoh model ruang vektor dengan dua dokumen.

Perhitungan dalam VSM didasarkan pada geometri di mana setiap istilah memiliki dimensi sendiri dalam ruang multi-dimensi, kueri dan dokumen adalah titik atau vektor dalam ruang ini. Ukuran cosine sering digunakan untuk menghitung ukuran kesamaan dan untuk menentukan sudut antara vektor dokumen dan vektor kueri. Rumus ukuran kosinus ini adalah:

$$similarity(\vec{d}_j, \vec{q}) = \frac{\vec{d}_j . \vec{q}}{|\vec{d}_j| |\vec{q}|} = \frac{\sum_{i=1}^{t} (w_{ij} . w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 . \sum_{i=1}^{t} w_{iq}^2}} \quad (1)$$

Dimana *d* adalah dokumen, *q* merupakan *query*, *t* adalah istilah, w_{ij} merupakan bobot Tf×Idf kata ke-*i* dari dokumen ke-*j*, dan w_{iq} merupakan bobot Tf×Idf kata ke-*i* dari query.

Bobot istilah dalam vektor dokumen dapat ditentukan menggunakan metode Tf × Idf. Bobot istilah diukur seberapa sering istilah *j* muncul dalam dokumen *i* (frekuensi frekuensi tf_i, *j*) dan di seluruh kumpulan dokumen (frekuensi dokumen df_j (jumlah dokumen yang mengandung istilah *j*)). Bobot istilah *j* dalam dokumen *i* adalah:

$$W_{dt} = tf_{dt} * IDF_t \quad (2)$$

Dimana *d* adalah dokumen ke-*d*, *t* merupakan istilah ke-*t* dari dokumen, W adalah notasi bobot dokumen ke-*d* terhadap istilah ke-*t*, sedangkan *tf* menunjukkan banyaknya istilah *i* pada sebuah dokumen**,** *idf* adalah singkatan dari *Inversed Document Frequency* diperoleh dengan *log2(n/df)***,** *n* adalah total dokumen dan *df* menunjukkan banyaknya dokumen dokumen yang mengandung istilah *i*.

Dalam menggunakan model ruang vektor untuk mengambil informasi, pengguna meminta basis data untuk menemukan dokumen yang relevan, dengan menggunakan representasi vektor dari dokumen yang terlibat. Penggunaan pencocokan kueri, dimungkinkan untuk menemukan dokumen yang paling mirip dengan permintaan yang digunakan dan juga untuk menentukan bobot istilah menggunakan ukuran cosinus. Misalkan untuk menggunakan kumpulan istilah dan dokumen, digambarkan proses pencocokan kueri berdasarkan ruang vektor enam dimensi menggunakan enam (6) istilah dan lima (5) dokumen. Berdasarkan proses pencarian kueri per dokumen, kami menulis kueri yang sesuai sebagai vektor menggunakan satu (1) untuk mewakili istilah kueri yang muncul di setiap dokumen dan nol (0) untuk mewakili *non-term* dalam setiap dokumen. Selanjutnya dicari dokumen yang relevan dengan menghitung sudut antara vektor kueri dan vektor dokumen. Nilai ambang 0,5 ditetapkan untuk menentukan apakah suatu dokumen akan dikembalikan relevan atau tidak. Penggunaan nilai ambang ini akan mengembalikan dokumen sebagai relevan hanya jika kosinus sudut yang dibuatnya dengan vektor kueri lebih besar dari nilai ambang yang telah ditetapkan. Dengan demikian semua dokumen yang memiliki istilah 1 (T1) dikembalikan

sebagai relevan. Menggunakan perbandingan kueri untuk menentukan dokumen yang paling relevan, kami menggunakan kosinus sudut antara vektor kueri dan vektor dokumen. Dengan menggunakan VSM, koleksi dokumen dapat diwakili oleh matriks bobot istilah, seperti contoh pada Gambar 3.

$$\begin{bmatrix} & T_1 & T_2 & ..... & T_t \\ D_1 & w_{11} & w_{21} & ..... & w_{t1} \\ D_2 & w_{21} & w_{22} & ..... & w_{t2} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ D_n & w_{1n} & w_{2n} & ..... & w_{tn} \end{bmatrix}$$

Gambar 3. Contoh matriks bobot istilah terhadap dokumen

3. Metode

3.1. Klasifikasi *Pretopology*

Prinsip utama dari klasifikasi dokumen adalah mencocokkan dokumen baru dengan struktur kelompok dokumen yang sudah diketahui sebelumnya. Algoritma klasifikasi *pretopology* yaitu mencocokkan dokumen baru terhadap dokumen representasi yang sudah terklasifikasi. Cara menentukan kelompok dokumen baru yaitu dengan mencari kemiripan *cosin similarity* terhadap dokumen representasi dari kelompok bidang. Dalam menentukan dokumen representasi dapat dilakukan dengan beberapa metode yaitu [5]:

1. Semua dokumen dalam suatu kategori.
2. Memilih 30 dokumen secara *random* dari suatu kategori.
3. Memilih 3 dokumen secara *random* dari suatu kategori.
4. Memilih 1 dokumen secara *random* dari suatu kategori.

Algoritma *pretopology* yang digunakan pada penelitian ini adalah algoritma "Kategorisasi Pretopologi Berdasarkan Perbandingan Kemiripan Dokumen" [10]. Dalam algoritma ini kategorisasi yang ada sebanyak *n* dilambangkan dengan $catSets_n$, sehingga nantinya akan ada *n* perwakilan dokumen, yang dilambangkan oleh $repDoc_i$, yang artinya setiap kategori memiliki satu perwakilan document yang mewakili kategorinya [10]. Dokumen yang akan diklasifikasi dilambangkan dengan d_j, dimana nilai *j* adalah 1 sampai m banyaknya dokumen yang baru. Setiap dokumen baru d_j akan dihitung nilai *cosine*-nya dengan perwakilan dokumen $repDoc_i$, dan menyimpannya dalam variable cos_{ij}. Sehingga nantinya untuk setiap dokumen baru, akan ada nilai kosinus n: cos_{1j}, cos_{2j}, ..., cos_{nj}. Dokumen yang memiliki nilai kosinus maksimal $max(cos_{nj})$ berarti dokumen yang diproses termasuk kedalam $catSets_n$.

Berikut algoritma dari klasifikasi *pretopology*:

1. Mendefinisikan *Docs[m]* sebagai daftar dokumen baru.
2. Mendefinisikan *repDocs[n]* sebagai dokumen representasi dari tiap – tiap kelas / kategori.
3. Mendefinisikan *catSets[n]* sebagai kelas.
4. Mendefinisikan *cos[n]* sebagai tempat menyimpan nilai kosinus antara *Docs[m]* dan *repDocs[n]*.
5. Menghitung nilai kosinus setiap dokumen baru *Docs[m]* dengan representasi dokumen *repDocs[n]* menggunakan fungsi *cosine similarity*.
6. Mencari nilai kosinus terbesar serta mengambil nilai n dan menetapkannya kedalam kategori *catSets[n]*.

3.2. Ekstraksi Kata Umum

Kata umum merupakan kata yang ruang lingkup maknanya lebih luas. Kata yang memiliki makna luas kurang tepat dalam proses pendefinisiannya. Pada penelitian ini, kata umum yang dimaksud adalah kata yang sering muncul pada bidang tertentu yang merupakan ekstraksi dari dukumen hasil klasifikasi. Pada penelitian [11], pendekatan *Zipf's Law* digunakan untuk menentukan *stopword* secara otomatis yaitu berdasarkan peringkat distribusi *term* frekuensi dalam koleksi dokumen. Sehingga asumsi yang digunakan adalah ekstraksi kata umum bidang ditentukan berdasarkan *term* frekuensi bidang yang kemudian dianggap sebagai *stoplist*.

```
Buat daftar term frekuensi dari
koleksi dokumen.
Urutkan term frekuensi secara
DESC, yaitu term frekuensi
tertinggi berada di paling atas.
Rangking daftar term frekuensi.
Yang memiliki frekuensi tertinggi
akan diset ranking = 1,
berikutnya ranking = 2 dan
seterusnya.
Gambarkan grafik dari term
frekuensi vs rangking. Hal ini
harus mematuhi Hukum Zipf's Law.
Pilih ambang batas dan setiap
kata yang muncul di atas ambang
batas tertentu dianggap sebagai
stopword.
Jalankan query dengan daftar
stopword tersebut, semua
stopwords dalam query harus
dihapus.
Evaluasi sistem setelah
menjalankan query dan catat rata-
rata presisi.
```

Gambar 4. Algoritma Pendekatan Dasar untuk Menentukan *Stopword* Otomatis

Algoritma pada Gambar 4 tersebut digunakan untuk menentukan kata umum bidang. Ambang batas yang digunakan adalah 1%, 2% dan 3% dengan mencari nilai terbaik hasil perbandingan dari ketiga ambang batas tersebut. Ambang batas dapat berubah sesuai kondisi atau data yang berbeda.

3.3. Evaluasi

Evaluasi digunakan untuk mengukur kinerja suatu sistem, dalam penelitian ini digunakan dua evaluasi meliputi evaluasi proses klasifikasi *pretopology* dan pencarian dokumen.

1. Evaluasi Proses Klasifikasi

Klasifikasi dokumen yang dievaluasi adalah nilai *recall* dan *precision*. Untuk mengukur keakuratan metode klasifikasi dokumen teks digunakan rumus sebagai berikut :

$$Recall = \frac{tp}{tp + fn} \quad (3)$$

$$Precision = \frac{tp}{tp + fp} \quad (4)$$

dimana *tp* (*True Positif*) adalah total dokumen yang terklasifikasi dengan benar, *fn* (*False Negative*) adalah total dokumen yang seharusnya merupakan kategori A ternyata terklasifikasi ke dalam ketegori yang lain, dan *fp* (*False Positive*) adalah dokumen dari kategori lain yang terklasifikasi ke dalam ketegori A.

2. Evaluasi Proses Pencarian

Evaluasi proses pencarian dilakukan untuk mengetahui tingkat relevansi dari proses pencarian dokumen. Pada penelitian ini, evaluasi proses pencarian dilakukan untuk mengukur keakuratan antara pencarian dokumen menggunakan proses penghapusan *stoplist* dengan pencarian

dokumen menggunakan penghapusan *stoplist* dan kata umum. Metode evaluasi yang biasanya digunakan untuk kasus penemuan kembali informasi atau dokumen adalah *recall* dan *precision* [12]–[14] yang dirumuskan sebagai berikut:

$$Recall = \frac{Document\ Relevant\ Items\ Retrieved}{Document\ Relevant\ Items} \quad (5)$$

$$Precision = \frac{Document\ Relevant\ Items\ Retrieved}{Document\ Items\ Retrieved} \quad (6)$$

Rata-rata dari nilai *precision* dan *recall* disebut dengan *F-Measure* yang dinyatakan pada persamaan berikut:

$$F\text{ - }measure = 2\frac{Precision\ x\ Recall}{Precision + Recall} \quad (7)$$

4. Hasil dan Pembahasan

Uji coba yang dilakukan terdiri dari dua yaitu klasifikasi dokumen dan pencarian dokumen. Pada proses pencarian dokumen fokus utamanya adalah membandingkan efektifitas antara aplikasi pencarian-1 dan aplikasi pencarian-2. Pada aplikasi pencarian-2 terdapat 3 tipe pencarian yaitu dengan *threshold* kata umum 1%, 2% dan 3%. Proses uji coba *query* dilakukan dengan 6 *query* terhadap 285 dokumen teknik, 193 dokumen ekonomi dan 268 dokumen pertanian. *Query* yang diujicobakan antara lain:

1. Sistem *fuzzy*;
2. Tanaman tembakau;
3. *Fuzzy;*
4. Teknologi mobile;
5. Sistem pendukung keputusan;
6. Manajemen analisis laba.

Query yang diujicobakan meliputi *query* yang tidak terdapat pada kata umum bidang dan kombinasi antara keduanya.

Analisa dilakukan untuk mengevaluasi kinerja dari metode dan sistem yang telah dibuat dan dilakukan uji coba. Analisa dan penelitian ini meliputi analisa terhadap proses klasifikasi dokumen yang menggunakan metode *pretopology* dan analisa terhadap proses pencarian dokumen pada aplikasi pencarian-1 dan aplikasi pencarian-2 dengan *threshold* kata umum bidang 1%, 2% dan 3%.

4.1. Analisa Klasifikasi

Setelah dilakukan penambahan dokumen sebanyak 25 dokumen teknik, 25 dokumen ekonomi dan 25 dokumen pertanian dan dilakukan proses klasifikasi kemudian analisa hasil klasifikasi, diperoleh hasil seperti pada Tabel 1.

Tabel 1. Tabel Hasil Analisa Proses Klasifikasi *Pretopology*

Bidang	Teknik	Ekonomi	Pertanian	Jumlah Dok	Recall	Precision
Teknik	31	6	8	45	0.69	0.57
Ekonomi	0	18	3	21	1	0.75
Pertanian	0	0	9	9	1	0.26
Rata-rata					0.90	0.76

Berdasarkan hasil uji coba proses klasifikasi diperoleh nilai rata-rata *recall* dan *precision* sebesar 90% dan 76%. Berdasarkan hasil analisa yang diperoleh maka dapat dikatakan bahwa metode *pretopology* cukup baik untuk proses klasifikasi dokumen.

4.2. Analisa Pencarian

Analisa pencarian ini dilakukan terhadap dua aplikasi pencarian, yaitu aplikasi pencarian-1 dan aplikasi pencarian-2. Analisa dilakukan untuk mengetahui efektifitas dari keduanya.

1. Aplikasi Pencarian-1

Setelah dilakukan uji coba *query* dengan 6 *query* seperti yang dijelaskan pada skenario di atas, untuk aplikasi pencarian-1 seperti pada Tabel 2, diperoleh nilai rata-rata *recall* = 96,7%.

precision = 22.5%, *f-measure* = 30.6% dan waktu proses 0,239 Detik.

Tabel 2. Tabel Hasil Analisa Proses Pencarian-1

No.	Query	Recall	Precision	F-Measure	Waktu Proses
1.	Sistem fuzzy	1	0,12	0,21	0,06
2.	Tanaman tembakau	1	0,11	0,19	0,05
3.	Fuzzy	1	0,83	0,91	0,02
4.	Teknologi mobile	1	0,11	0,19	0,29
5.	System pendukung keputusan	1	0,15	0,26	0,56
6.	Manajemen analisis data	0,8	0,03	0,06	0,47
Rata-rata		0,97	0,23	0,04	0,24

2. Aplikasi Pencarian-2 dengan *Threshold* Kata Umum 1%

Dengan *query* yang sama pada uji coba aplikasi pencarian-1, untuk aplikasi pencarian-2 dengan *threshold* 1% yang dapat dilihat pada Tabel 3 diperoleh nilai rata-rata *recall, precision, f-measure,* dan waktu proses berturut-turut adalah 96%, 67.3%, 76.5% dan 0,098 detik.

Tabel 3. Tabel Hasil Analisa Proses Pencarian-2 dengan Threshold Kata Umum 1%

No.	Query	Recall	Precision	F-Measure	Waktu Proses
1.	Sistem fuzzy	1	0,96	0,98	0,02
2.	Tanaman tembakau	1	0,81	0,89	0,02
3.	Fuzzy	1	0,83	0,91	0,02
4.	Teknologi mobile	1	0,78	0,88	0,13
5.	System pendukung keputusan	0,96	0,48	0,64	0,23
6.	Manajemen analisis data	0,8	0,18	0,29	0,16
Rata-rata		0,96	0,67	0,77	0,10

3. Aplikasi Pencarian-2 dengan *Threshold* Kata Umum 2%

Dengan *query* yang sama pula pada uji coba aplikasi pencarian-1, Tabel 4 untuk aplikasi pencarian-2 dengan *threshold* 2% diperoleh nilai rata-rata *recall, precision, f-measure* dan waktu proses berturut-turut adalah 63.3%, 51.2%, 56.3% dan 0,045 Detik.

Tabel 4. Tabel Hasil Analisa Proses Pencarian-2 dengan Threshold Kata Umum 2%

No.	Query	Recall	Precision	F-Measure	Waktu Proses
1.	Sistem fuzzy	1	0,96	0,98	0,02
2.	Tanaman tembakau	0	0	0	-
3.	Fuzzy	1	0,83	0,91	0,02
4.	Teknologi mobile	1	0,78	0,88	0,12
5.	System pendukung keputusan	0	0	0	-
6.	Manajemen analisis data	0,8	0,50	0,62	0,11
Rata-rata		0,63	0,51	0,56	0,05

Keterangan:
(-) = Dokumen tidak ditemukan

Tabel 5. Tabel Hasil Analisa Proses Pencarian-2 dengan Threshold Kata Umum 3%

No.	Query	Recall	Precision	F-Measure	Waktu Proses
1.	Sistem fuzzy	0	0	0	-
2.	Tanaman tembakau	0	0	0	-
3.	Fuzzy	0	0	0	-
4.	Teknologi mobile	1	0,78	0,88	0,12
5.	System pendukung keputusan	0	0	0	-
6.	Manajemen analisis data	0,8	0,50	0,62	0,12
Rata-rata		0,30	0,21	0,25	0,06

Keterangan:
(-) = Dokumen tidak ditemukan

4. Aplikasi Pencarian-2 dengan *Threshold* Kata Umum 3%

Tabel 5 menunjukkan hasil pencarian dengan *query* yang sama pula pada uji coba aplikasi Pencarian-1, untuk aplikasi pencarian-2 dengan *threshold* 3% diperoleh nilai rata-rata *recall, precision. f-measure* dan waktu proses berturut-turut adalah 30%, 21.3%, 24.9% dan 0,061 Detik.

Dari perbandingan *threshold* kata umum tersebut diperoleh aplikasi pencarian-2 terbaik adalah dengan *threshold 1%.* Aplikasi pencarian-2 dengan *threshold* 1% kemudian dibandingkan dengan aplikasi pencarian-1. Adapun hasil perbandingan dari keduanya tercantum pada Tabel 6.

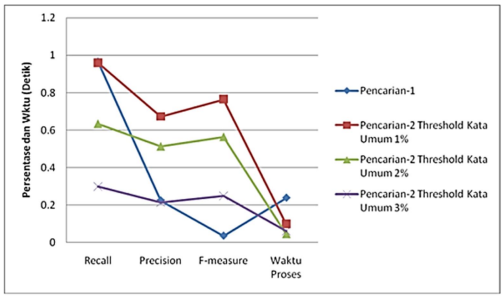
Tabel 6. Tabel Perbandingan Hasil Analisa Proses Pencarian-1 dan Pencarian-2 dengan *Threshold* Kata Umum 1%

No.	Aplikasi	Recall	Precision	F-Measure	Waktu Proses
1.	Pencarian-1	0,97	0,23	0,31	0,24
2.	Pencarian-2	0,96	0,67	0,77	0,10

Dari perbandingan tersebut diperoleh nilai *recall, precision, f-measure* dan waktu proses untuk aplikasi pencarian-1 adalah 96,7%, 22,5%, 30,6% dan 0,239 Detik. Sedangkan untuk aplikasi pencarian-2 dengan threshold 1% adalah 96%, 67,3%, 76,5% dan 0,098 Detik.

Hasil analisa perbedaan *recall*, *precision*, *f-measure* dan waktu proses keempat tipe pencarian tersebut dapat digambarkan dalam bentuk grafik perbandingan sebagaimana ditunjukkan pada Gambar 5.

Dari grafik tersebut, terlihat bahwa pencarian yang paling efektif dengan waktu proses paling cepat adalah tipe pencarian-2 dengan threshold kata umum 1%.



Gambar 5. Grafik Perbandingan Analisa Pencarian-1, Pencarin-2 dengan *Threshold* Kata Umum 1%, 2% dan 3%

5. Penutup

Berdasarkan hasil uji coba dan evaluasi sistem, maka dapat disimpulkan sebagai berikut:

1. Klasifikasi 25 dokumen teknik, 25 ekonomi dan 25 pertanian dengan menerapkan metode *Pretopology* menghasilkan nilai rata-rata *recall* dan *precision* sebesar 90% dan 76%.
2. Ekstraksi kata umum bidang menggunakan pendekatan *Zipf's Law* dengan melakukan perbandingan *threshold* 1%, 2% dan 3%. Hasil uji coba diperoleh *threshold* kata umum terbaik adalah 1% terhadap 285 dokumen teknik, 193 ekonomi dan 268 pertanian.
3. Aplikasi pencarian dengan 6 *query* terhadap 746 dokumen yang terklasifikasi menjadi: 285 dokumen teknik, 193 ekonomi dan 268 pertanian untuk aplikasi pencarian-1 mampu menghasilkan nilai rata-rata *recall, precision, f-measure* dan waktu proses sebesar 96,7%, 22,5%, 30,6% dan 0.239 Detik. Sedangkan aplikasi pencarian-2 berturut-turut adalah 96%, 67.3%, 76.5% dan 0.098 Detik. Sehingga dapat dikatakan bahwa sistem pencarian-2 lebih efektif dari pada sistem pencarian-1.

Saran untuk penelitian selanjutnya, pada ekstraksi kata umum

diperoleh beberapa kata umum yang seharusnya tidak termasuk dalam kata umum bidang dan perlu dilakukan penelitian dengan jumlah dokumen yang lebih banyak.

6. Daftar Pustaka

[1] D. B. Deshmukh and Y. Pandey, "A Review on Hierarchical Document Clustering," *J. Data Min. Knowl. Discov.*, vol. 3, no. 5, pp. 65–68, 2012.

[2] F. A. Hermawati and D. A. Zuhdi, "Aplikasi Sistem Temu Kembali Dokumen dengan Metode Vector Space Model," *KONVERGENSI*, vol. 5, no. 2, pp. 38–49, 2009.

[3] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," 2003.

[4] K. J. Cios, W. Pedrycz, R. W. Swiniarski, and L. A. Kurgan, *Data Mining : A Knowledge Discovery Approach*. Springer, 2007.

[5] C. D. Manning, P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval*. Cambridge, England: Cambridge University Press, 2009.

[6] A. Z. Arifin, I. P. A. K. Mahendra, and H. T. Ciptaningtyas, "Enhanced Confix Stripping Stemmer and ANTS Algorithm for Classifying News Docements in Indonesian Language," in *The 5th International Conference on Information & Communication Technology and Systems*, 2009, pp. 149–158.

[7] G. Tsatsaronis and V. Panagiotopoulou, "A generalized vector space model for text retrieval based on semantic relatedness," *EACL 2009 - 12th Conf. Eur. Chapter Assoc. Comput. Linguist. Proc.*, no. April, pp. 70–78, 2009.

[8] J. N. Singh and S. K. Dwivedi, "Performance Analysis of Layered Vector Space Model in Web Information Retrieval," *Int. J. Appl. Inf. Syst.*, vol. 8, no. 5, pp. 7–15, 2015.

[9] P. Harcourt and R. B. Japheth, "Application of Vector Space Model to Query Ranking and Information Retrieval," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 6, no. 5, pp. 42–47, 2016.

[10] M. Ahat, S. Amor, and M. Bui, "Document Classification with LSA and Pretopology.," *Stud. Inform. Universalis*, vol. 8, no. 1, pp. 125–144, 2010.

[11] R. T.-W. Lo, B. He, and I. Ounis, "Automatically Building a Stopword List for an Information Retrieval System," *J. Digit. Inf. Manag. Spec. Issue 5th Dutch-belgian Inf. Retr. Work.*, vol. 3, pp. 3–8, 2005.

[12] F. A. Hermawati, H. Tjandrasa, and N. Suciati, "Sistem Retrieval Citra Berbasis Region Dengan Transformasi Wavelet Berdasarkan Karakteristik Color-Texture," *KONVERGENSI*, vol. 2, no. 1, pp. 1–9, 2006.

[13] F. A. Hermawati, H. H. Tjandrasa, and N. Suciati, "Evaluasi Representasi Warna Untuk Retrieval Citra Berbasis Region," *J. Saintek*, vol. 9, no. 2, pp. 101–107, 2005.

[14] F. A. Hermawati, "Sistem Temu Kembali Citra Berdasarkan Karakteristik Bentuk dengan Metode Color-Edge Extraction," in *Seminar Nasional Teknik 2009*, 2009, pp. 253–257.