Information Retrieval pada Proses Penyimpanan dan Pencarian Dokumen Digital Menggunakan Metode *Text Mining*

Enda Esyudha Pratama

Fakultas Teknik Jurusan Informatika
Universitas Tanjungpura
enda@informatika.untan.ac.id
Jl. Prof. Dr. H. Hadari Nawawi, Pontianak 78124

Abstrak

Information Retrival (IR) atau temu balik informasi merupakan sebuah cara untuk mengorganisasikan, merepresentasikan, menyimpan, dan melakukan pencarian informasi dalam bentuk teks dan multimedia. Proses penyimpanan dan pencarian dokumen menjadi suatu hal yang sangat penting untuk pengelolaan dokumen digital dalam jumlah yang besar pada sebuah sistem. Sistem yang dibangun akan menerapkan metode *text mining* pada proses penyimpanan dan pencarian dokumen. Metode text mining digunakan untuk menggali informasi penting dari sebuah dokumen untuk kemudian menyimpannya sebagai index, tags, atau keyword dari sebuah dokumen dan melakukan proses kategorisasi secara otomatis. Informasi penting tersebut nantinya akan dicocokan dengan kata kunci pencarian untuk selanjutnya menampilkan dokumen hasil pencarian dan mengurutkannya berdasarkan tingkat relevansinya. Hasil pengujian yang dilakukan pada 50 dokumen untuk 5 kategori menunjukan fungsionalitas berjalan baik dan waktu proses untuk setiap dokumen rara-rata selama 0,0623detik.

Kata kunci: information retrieval, penyimpanan, pencarian, *text mining*

1 PENDAHULUAN

Pemanfaatan teknologi informasi di zaman sekarang ini telah banyak digunakan di berbagai bidang, tak terkecuali pengelolaan dokumen. Akibatnya jumlah dokumen digital yang sangat besar pun tak dapat dihindari. Sebagai contoh pada tingkat universitas yang umumnya mewajibkan setiap mahasiswanya menghasilkan skripsi. Bentuk skripsi yang dihasilkan biasanya tersedia dalam format digital dan disimpan dalam suatu sistem.

Sistem pencarian informasi maupun dokumen pada umumnya menampilkan hasil pencarian dalam daftar yang panjang. Kemudian pengguna diharuskan memilah sendiri dokumen mana yang relevan dengan topik yang mereka cari dalam daftar tersebut. Sayangnya sebagian besar *search engine* masih menggunakan paradigma tersebut. Kelemahan search engine tersebut membuat pengguna cukup kesulitan untuk menemukan informasi dari isi dokumen yang mereka cari (Büttcher, 2016).

Semakin besar jumlah koleksi dokumen yang dimiliki maka kecepatan dan ketepatan perolehan informasi sangat penting bagi pencari informasi. Salah satu cara yang paling banyak digunakan adalah *Information Retrieval* (IR) atau temu balik informasi.

Information Retrival (IR) atau temu balik informasi merupakan sebuah cara untuk mengorganisasikan, merepresentasikan, menyimpan, dan melakukan pencarian informasi dalam bentuk teks dan multimedia (Kowalsi, 1997). Terdapat beberapa metode yang dapat digunakan dalam membangun Information Retrieval System ini salah satunya adalah text mining.

Metode text mining dapat digunakan untuk memperoleh informasi penting dari suatu dokumen berdasarkan kata-kata yang ada dalam dokumen tersebut (Han & Kamber, 2006). Pada penelitian ini, akan digunakan metode text mining untuk mengolah data dari kandungan teks atau kata dari setiap dokumen yang diunggah ke dalam sistem dan memanggilnya kembali berdasarkan frekuensi kemunculan kata kunci pencarian di dalam teks dari isi dokumen tersebut.

Pada saat dokumen disimpan, dokumen akan dikategorikan secara otomatis sesuai dengan isi dari dokumen tersebut yang diperoleh dengan menggunakan metode text mining. Informasi penting dari kata-kata yang terdapat dalam dokumen tersebut diambil untuk disimpan sebagai tags atau keyword dalam database untuk nantinya dicocokan dengan kata kunci pencarian. Kemudian pada saat pengguna memasukkan kata kunci pencarian maka hal ini akan menjadi masukan untuk mencari dokumen yang relevan.

Berdasarkan uraian di atas maka akan dibangun Information Retrieval System pada Proses Penyimpanan dan Pencarian Dokumen Digital Menggunakan Metode Text Mining. Dengan adanya penelitian ini diharapkan kumpulan dokumen digital dapat dikelola secara baik dan dapat dengan mudah untuk disimpan dan dicari kembali secara cepat dan tepat sesuai dengan kebutuhan pengguna sistem.

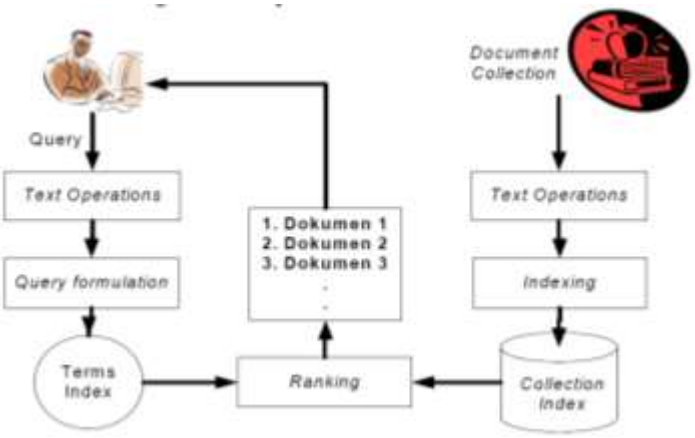
2 METODOLOGI PENELITIAN

2.1 *Information Retrieval*

Information Retrieval (IR) merupakan suatu cara yang digunakan untuk menemukan kembali (retrieve) informasi-informasi yang relevan terhadap kebutuhan pengguna dari suatu kumpulan informasi secara otomatis. (Bunyamin, 2015)

Proses IR umumnya berkaitan dengan pencarian informasi yang isinya tidak terstruktur. Demikian pula kata kunci pencarian pengguna yang disebut query, juga merupakan bentuk yang tidak struktur. Hal ini yang membedakan IR dengan sistem basis data. Dokumen adalah contoh informasi yang tidak terstruktur. Isi dari suatu dokumen umunya berupa kumpulan teks yang sangat tergantung pada pembuat dokumen tersebut (Pardede, 2014).

Secara umum, sistem IR memiliki beberapa bagian yang membangun sistem secara keseluruhan. Gambaran bagian-bagian yang terdapat pada suatu sistem IR dapat dilihat pada Gambar 1



Gambar 1 Arsitektur Information Retrieval (IR) System
Sumber: (Bunyamin, 2015)

Gambar 1 memperlihatkan bahwa terdapat dua buah alur operasi pada IR system. Alur pertama dimulai dari koleksi dokumen dan alur kedua dimulai dari query pengguna. Alur pertama yaitu pemrosesan terhadap koleksi dokumen menjadi basis data indeks tidak tergantung pada alur kedua. Sedangkan alur kedua tergantung dari keberadaan basis data indeks yang dihasilkan pada alur pertama.

Bagian-bagian dari IR system menurut gambar 2 meliputi (Bunyamin, 2015):

1. *Text Operations* (operasi terhadap teks) yang meliputi pemilihan kata-kata dalam query maupun dokumen (term selection) dalam pentransformasian dokumen atau query menjadi term index (indeks dari kata-kata).
2. *Query formulation* (formulasi terhadap query) yaitu memberi bobot pada indeks katakata query.
3. *Ranking* (perangkingan), mencari dokumen-dokumen yang relevan terhadap query dan mengurutkan dokumen tersebut berdasarkan kesesuaiannya dengan query.
4. *Indexing* (pengindeksan), membangun basis data indeks dari koleksi dokumen. Dilakukan terlebih dahulu sebelum pencarian dokumen dilakukan.

Salah satu aplikasi umum dari pemanfaatan IR adalah search engine atau mesin pencarian. Pengguna dapat mencari informasi apa saja di internet yang dibutuhkannya melalui search engine. Contoh lainnya dari pemanfaatan IR adalah sistem informasi perpustakaan yang digunakan untuk megelola proses penyimpanan dan pencarian dokumen digital.

2.2 *Text Mining*

Text mining adalah sebuah teknik/pendekatan algoritmik berbasis komputer untuk mendapatkan suatu pengetahuan baru yang tersembunyi dari sekumpulan teks. Text mining merupakan bagian dari keilmuan information retrieval (temu balik informasi) yang bekerja pada data bertipe teks yang cenderung tidak terstruktur (Priyanto, 2018).

Tahapan dalam text mining secara umum adalah tokenizing, filtering, stemming, tagging, dan analyzing (Prilianti, 2014). Tokenizing merupakan tahapan untuk memisah-misahkan setiap kata (token) pada dokumen input. Filtering merupakan proses seleksi terhadap kata-kata yang dihasilkan dari proses tokenizing, dapat dilakukan dengan algoritma stop list maupun word list. Algoritma stop list akan membuang kata-kata yang tidak penting seperti kata ganti, kata keterangan, kata sambung, kata depan dan kata sandang. Sebaliknya, algoritma word list akan menyimpan kata-kata yang penting. Proses stemming kemudian

dilakukan untuk mencari kata dasar dari setiap kata yang telah lolos proses filtering.

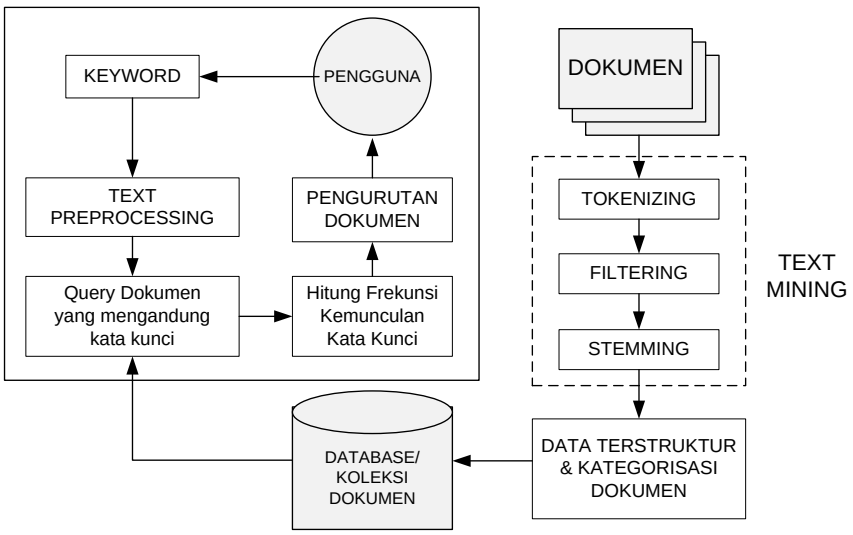
Pada dokumen yang berbahasa Indonesia, proses filtering secara sederhana dilakukan dengan menghilangkan awalan dan akhiran dari setiap kata. Jika dokumen berbahasa Inggris, maka diperlukan proses lanjutan yang disebut sebagai tagging. Proses tagging dilakukan untuk mencari bentuk awal dari setiap kata lampau. Setelah semua kata penting berhasil dikoleksi dari rangkaian proses tersebut, maka tahap berikutnya adalah analyzing yaitu menentukan keterhubungan antar dokumen dengan mengamati frekuensi kemunculan tiap kata yang ada pada tiap dokumen.

Pada penelitian ini, pendekatan yang digunakan adalah Term Frequency (TF) untuk menghitung frekeunsi kemunculan kata. Pendekatan TF tidak mengindahkan term yang terkandung dalam dokumen lain. Metode TF hanya secara sederhana menghitung kemunculan term dalam suatu dokumen. Term-term yang memiliki frekuensi kemunculan tinggi akan menjadi ciri dari suatu dokumen dimana term tersebut berada (Priyanto, 2018).

2.3 Rancangan Sistem

Bahan penelitian yang digunakan berupa kumpulan dokumen ilmiah. Kumpulan dokumen tersebut didapat dari laporan tugas akhir mahasiswa Univeristas Tanjungpura, jurnal ilmiah, karya tulis, dan makalah atau tugas kuliah mahasiswa.

Secara garis besar, sistem terdiri dari dua proses utama yaitu proses preprocessing dokumen menggunakan text mining dan proses pencarian untuk menemukan dokumen yang memiliki keterkaitan dengan kata kunci atau query dari pengguna. Proses text preprocessing isi dokumen menggunakan text mining dilakukan untuk mengubah isi dari dokumen yang memiliki sifat tidak terstruktur menjadi data tersturktur. Data terstruktur tersebut merupakan informasi penting yang terkandung berdasarkan kumpulan kata atau teks dari isi dokumen yang nantinya akan digunakan sebagai sumber data untuk proses selanjutnya. Arsitektur sistem dapat dilihat pada Gambar 2.



Gambar 2 Rancangan Arsitektur Sistem

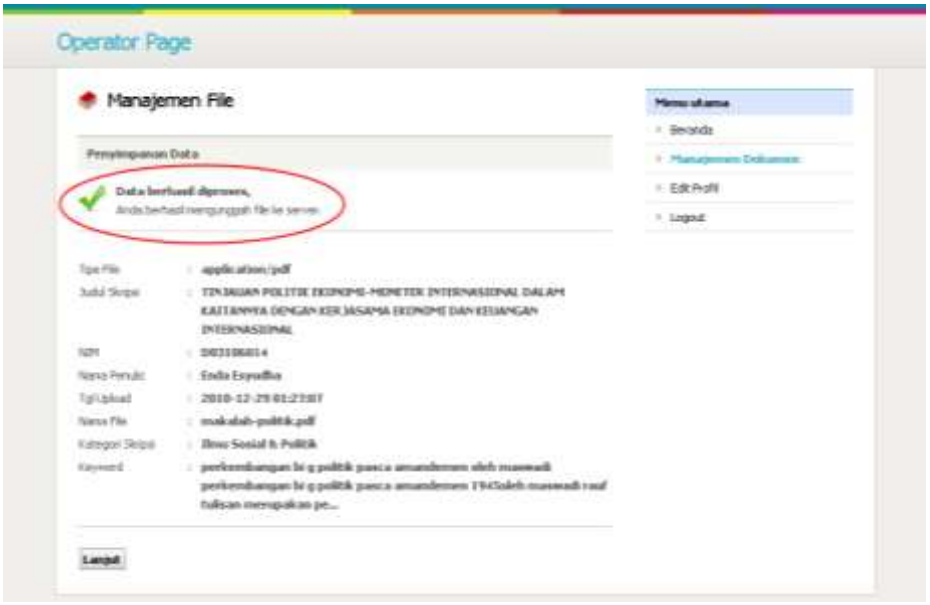
Proses pencarian merupakan proses menemukan kembali informasi (dokumen) yang

relevan terhadap query kata kunci yang diberikan oleh pengguna dan menyediakan daftar dokumen terurut yang relevan dengan query tersebut. Tahapan ini dimulai dengan menerima query kata kunci dari pengguna.

Kemudian dilakukan tahap *text preprocessing* pada kata kunci tersebut. Kata kunci akan dibandingkan dengan kumpulan kata atau teks dari isi dokumen-dokumen yang telah diproses sebelumnya pada saat dokumen diunggah. Keterkaitan antara kata kunci dengan dokumen dihitung berdasarkan frekuensi kemunculan kata kunci dalam isi dokumen tersebut. Semakin besar frekunsi kemunculan kata kunci, maka semakin besar keterkaitan antara dokumen dengan kata kunci. Daftar dokumen yang memiliki keterkaitan akan diurutkan berdasarkan frekunsi kemunculannya. Daftar terurut inilah yang akan dikembalikan kepada pengguna..

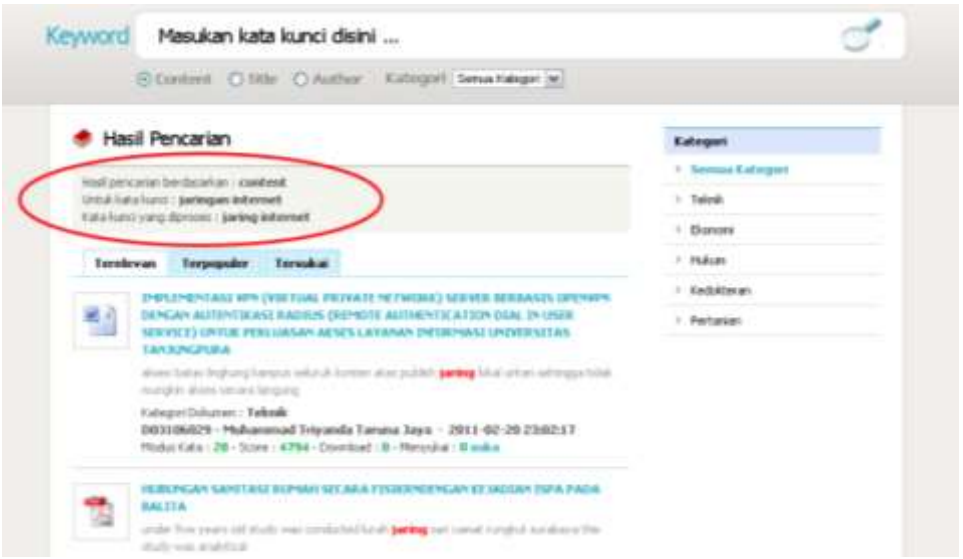
3 HASIL DAN PEMBAHASAN

Berdasarkan hasil perancangan yang telah dilakukan, dibuatlah suatu sistem yang dapat mengelola proses penyimpanan dan pencarian dokumen digital. Proses information retrieval dengan menggunakan metode text mining dilakukan pada saat dokumen ditambahkan dan disimpan ke dalam sistem mapun saat dicari kembali. Adapun proses penambahan data dokumen dapat dilihat pada Gambar 3.



Gambar 3 Proses Penyimpanan Dokumen

Pada gambar tersebut dapat dilihat, ketika proses penyimpanan berhasil akan ditampilkan sejumlah keyword hasil text mining dari isi dokumen yang telah diproses secara otomatis. Selain itu, proses kategorisasi juga dilakukan secara otomatis. Selanjutnya pada proses pencarian, pengguna diminta untuk memberikan keyword yang berkaitan dengan dokumen yang ingin dicarinya. Proses pencarian pada sistem dapat dilihat pada Gambar 4.



Gambar 4 Tampilan Hasil Pencarian

Pada gambar tersebut diperlihatkan bagian dari panel informasi hasil proses pencarian. Informasi yang ditampilkan berupa kata kunci dan parameter pencarian. Kemudian di bagian bawah dari panel informasi tersebut ditampilkan daftar dokumen yang mempunyai relevansi dengan kata kunci pencarian.

Pada proses pengujian, metode yang digunakan adalah dengan menggunakan 10 dokumen dari 5 kategori sehingga total dokumen yang digunakan berjumlah 50 dokumen. Dokumen yang digunakan memiliki variasi jumlah 1.000-3.000 kata. Berdasarkan pengujian yang dilakukan didapat waktu rata-rata untuk memproses tiap satu dokumen yaitu sekitar 0,0623 detik. Hasil perhitungan waktu proses dapat dilihat pada Tabel 1.

Tabel 1 Tabel Perhitungan Waktu Proses Tiap Kata

No	Teknik	Ekonomi	Hukum	Pertanian	Kedokteran
	Waktu (detik)	Waktu (detik)	Waktu (detik)	Waktu (detik)	Waktu (detik)
1	0,1142	0,0528	0,0693	0,0619	0,1018
2	0,0793	0,0736	0,0998	0,0456	0,0125
3	0,0885	0,0727	0,0591	0,0832	0,0632
4	0,0683	0,0732	0,0643	0,0409	0,0723
5	0,0789	0,0671	0,0600	0,0528	0,0536
6	0,0743	0,0548	0,0608	0,0559	0,0810
7	0,0628	0,0521	0,0689	0,0543	0,0652
8	0,0552	0,0505	0,0698	0,0481	0,0449
9	0,0542	0,0423	0,0702	0,0391	0,0559
10	0,0546	0,0521	0,0516	0,0408	0,0437
	0,0730	**0,0591**	**0,0676**	**0,0523**	**0,0594**
	Rata-Rata = 0,0623 detik				

4 KESIMPULAN

Berdasarkan hasil perancangan dan implementasi memperlihatkan sistem dapat melakukan information retrieval pada proses penyimpanan dan pencarian dokumen dengan metode text mining. Pengujian blackbox menunjukan fungsionalitas sistem berjalan baik. Sedangkan rata-rata waktu proses yang dibutuhkan untuk memproses tiap dokumen sebesar 0,0623 detik.

Referensi

Büttcher, S., Clarke, C. L., & Cormack, G. V. (2016). Information retrieval: Implementing and evaluating search engines. Mit Press.

Kowalski, G. (1997). Information Retrieval System Theory and Implementation, Kluwer Academic Publisher, United States of America.

Han, J., Kamber, M. (2006) Data Mining Concept and Technique, 2nd Ed, Elsevier.

Bunyamin, H. (2015). 8. Algoritma Umum Pencarian Informasi Dalam Sistem Temu Kembali Informasi Berbasis Metode Vektorisasi Kata dan Dokumen. Jurnal Informatika, 1(2).

Pardede, J. (2014). Implementasi Multithreading Untuk Meningkatkan Kinerja Information Retrieval Dengan Metode GVSM. Jurnal Sistem Komputer, 4(1), 1-6.

Priyanto, A., & Ma'arif, M. R. (2018). Implementasi Web Scrapping dan Text Mining untuk Akuisisi dan Kategorisasi Informasi dari Internet (Studi Kasus: Tutorial Hidroponik). Indonesian Journal of Information System, 1(1), 25-33.

Prilianti, K. R., & Wijaya, H. (2014). Aplikasi Text Mining untuk Automasi Penentuan Tren Topik Skripsi dengan Metode K-Means Clustering. Jurnal Cybermatika, 2(1)..