STMIK DCI

JURNAL TEKNIK INFORMATIKA

Halaman Jurnal: http://jurnal.stmik-dci.ac.id/index.php/jutekin/

Halaman LPPM STMIK DCI : http://lppm.stmik-dci.ac.id/



PENERAPAN MESIN PENCARI INFORMASI DENGAN MENGGUNAKAN METODE VECTOR SPACE MODEL

Ridwan[1], Tomi Alfian Armawan Sandi[2]

Ilmu Komputer, STMIK Nusa Mandiri

Email : ridwans70@gmail.com[1] , alfian.armawan@gmail.com[2]

ABSTRAK

Pertumbuhan pencari mesin saat ini menghasilkan tingkat penarikan yang tinggi dan tingkat presisi yang rendah. Ingat mana yang tinggi untuk ditafsirkan bahwa dokumen yang dihasilkan dalam penelusuran dokumen banyak, sedangkan tingkat akurasi rendah dapat diartikan bahwa dokumen yang diharapkan dapat ditemukan sedikit / sedikit atau lebih rendah. Solusi untuk mengatasi masalah di atas dengan membuat sistem informasi pengembalian pertemuan menggunakan metode Vector Space Model (VSM). Metode VSM dipilih karena cara pemodelan yang efisien, mudah dalam representasi dan mendapatkan implementasi pada pencocokan dokumen.

Kata Kunci : Mesin, Informasi, Vector, Penerapan, Space, Model.

I. PENDAHULUAN

Informasi merupakan dapat dikatakan sebagai aset yang sangat bernilai. Apalagi Jika informasi tersebut dapat dikelola dengan baik dan bijak. Informasi bisa mudah didapatkan karena perkembangan teknologi informasi yang semakin cepat dan pengetahuan setiap orang yang terus bertambah. Teknologi informasi khususnya internet sangat mendukung terjadinya pertukaran informasi dengan sangat cepat. Internet menjadi media informasi dan telekomunikasi yang telah dimanfaatkan banyak orang dengan banyak kepentingan. Internet menciptakan banyak informasi dan karena kemudahan dalam akses internet inilah banyak orang menggunakan media online dalam kegiatannya sehari-hari. Setiap orang membutuhkan informasi dan mereka bisa mencari informasi dengan menggunakan mesin pencari yang ada didunia maya.

Penelitian terkait dengan metode vector space model juga dilakukan antara lain Dr. Khalaf Khatatneh, M. Wedyan, DR. Mohamed Alham, DR Basem Alrifai (2005) dalam publikasinya yang berjudul ”Using new Data Structure to Implement Documents Vectors in Vector Space Model in Information Retrieval System” Bagaimana menggunakan tabel terstruktur dan vector space model untuk lebih menghemat ruang untuk file dokumen yang sebelumnya memerlukan space besar dalam sistem temu kembali informasi. Aplikasinya dilakukan dengan menggunakan tabel dimana pada baris pertama digunakan kata kunci dan baris kedua digunakan pembobotan dari setiap kata kunci.

Solusi untuk mengatasi masalah di atas adalah dengan membuat sistem temu kembali informasi menggunakan metode Vector Space Model (VSM). Metode VSM dipilih karena cara kerja model ini efisien, mudah dalam representasi dan dapat diimplementasikan pada document-matching.

II. LANDASAN TEORI

2.1 Sistem Temu Kembali Informasi

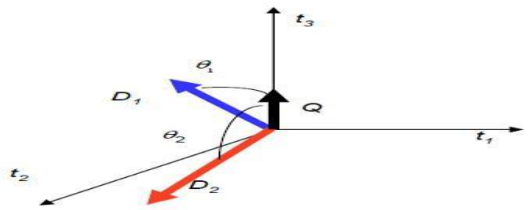
Sistem temu kembali informasi merupakan bagian dari ilmu komputer yang berhubungan dengan pengambilan informasi dari dokumen-dokumen yang didasarkan pada isi dan konteks dari dokumen-dokumen itu sendiri. Proses dalam sistem temu kembali dapat digambarkan sebagai sebuah proses untuk mendapatkan dokumen yang relevan dari koleksi dokumen melalaui pencarian query yang diimputkan user

2.2 Text Mining

Text mining adalah salah satu bidang khusus dari data mining. Sesuai buku The Text Mining Handbook, text mining dapat didefinisikan sebagai suatu proses menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan tools analisis yang merupakan komponen-komponen dalam data mining yang salah satunya adalah kategorisasi. Tujuan dari text mining adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Salah satu elemen kunci dari text mining adalah kumpulan dokumen yang berbasis teks. Pada prakteknya, text mining ditujukan untuk menemukan pola dari sekumpulan dokumen yang jumlahnya sangat besar dan bisa mencapai jumlah ribuan bahkan sampai jutaan. Koleksi dokumen bisa statis, dimana dokumen tidak berubah, atau dinamis, dimana dokumen selalu diupdate sepanjang waktu.

1. Vector Space Model

Vector Space Model (VSM) adalah metode untuk melihat tingkat kedekatan atau kesamaan (smilarity) term dengan cara pembobotan term. Dokumen dipandang sebagi sebuah vektor yang memiliki magnitude (jarak) dan direction (arah). Pada Vector Space Model, sebuah istilah direpresentasikan dengan sebuah dimensi dari ruang vektor. Relevansi sebuah dokumen ke sebuah query didasarkan pada similaritas diantara vektor dokumen dan vektor query.



Gambar 1. Ilustrasi Vector Space Model

Dimana:
ti = Kata di database
Di = Dokumen
Q = Kata Kunci

Cara kerja dari vector space model adalah dengan menghitung nilai cosines sudut dari dua vector, yaitu vektor kata kunci terhadap vektor tiap dokumen. Perhitungan vektor space model menggunakan persamaan (1),(2) dan (3)

$$Cosine\ \theta_{D_i} - Sim(Q, D_i)$$

Dimana
Q = query (kata kunci)
Di = dokumen ke-i

$$Sim(Q, D_i) = \frac{\sum_j W_{i,j}\ W_{q,j}}{\sqrt{\sum_j W_{i,j}^2}\sqrt{\sum_i W_{q,j}^2}}$$

Dimana:
Di = dokumen ke-i
Q = query (kata kunci)
J = Kata diseluruh dokumen

$$Cosine\ \theta_{D_i} = \frac{Q \cdot D_i}{|Q|}$$

Dimana:
Di = dokumen ke-i
Q = query (kata kunci)
[Q] = Vektor Q
[Di] = Vektor Di

III. ANALISIS MASALAH

Berdasarkan latar belakang di atas, maka permasalahan yang dapat dirumuskan adalah bagaimana membuat aplikasi dengan algoritma yang efektif dalam pencarian dokumen dengan metode vec

tor space model pada dokumen teks berbahasa Indonesia, sehingga pengguna mudah mencari dokumen yang diinginkan. Dalam penelitian ini ada beberapa pembatasan masalah yang dilakukan, yaitu:

a. Dokumen yang digunakan adalah dokumen teks berbahasa Indonesia.
b. Penelitian ini digunakan metode Vector Space Model

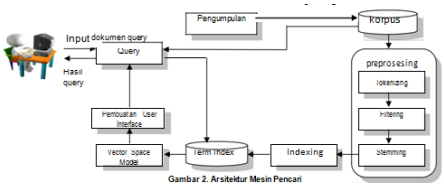
IV. PERANCANGAN SISTEM

4.1 Arsitektur Sistem Temu Kembali

Information Retrieval atau sistem temu kembali dengan menggunakan Vector Space Model sebagai suatu sistem memiliki beberapa proses (modul) yang membangun system secara keseluruhan. Modul system temu kembali informasi terdiri dari : modul pengumpulan dokumen, modul tokenisasi (tokenizing), modul

pembuangan stopword (filtering), modul Pengubahan kata dasar (stemming), modul Pengindeksan kata (indexing), modul Vector Space Model (term similarity) dan modul pembuatan user interface.

Arsitekturnya dapat dilihat pada gambar 2 dibawah ini.



Gambar 2. Arsitektur Mesin Pencari

Gambar 4.1

Arsitektur Mesin Pencari

4.2 Implementasi Vector Space Model

Modul sistem temu kembali informasi terdiri dari : modul pengumpulan dokumen, modul tokenisasi (tokenizing), modul pembuangan stopword (filtering), modul Pengubahan kata dasar (stemming), modul Pengindeksan kata (indexing), dan modul Vector Space Model (term similarity). Modul Pengumpulan Dokumen. Proses pengumpulan dokumen-dokumen yang dipilih untuk disimpan dalam korpus. Dokumen-dokumen yang dipilih adalah dokumen teks berbahasa Indonesia. Pada penelitian ini digunakan 500 abstraksi skripsi mahasiswa Universitas Indonesia.

Modul Tokenisasi (tokenizing). Tokenizing adalah proses pemotongan string input berdasarkan tiap kata yang menyusunnya. Umumnya setiap kata yang teridentifikasi atau terpisahkan dengan kata yang lain oleh karakter spasi, sehingga proses ini menggunakan spasi pada dokumen untuk melakukan pemisahan kata. Modul Pembuangan stopword (filtering). Tahap filtering adalah proses pembuangan term yang tidak memiliki arti atau tidak relevan. Term yang diperoleh pada tahap tokenisasi dicek dalam suatu daftar stopword, jika term masuk dalam daftar stopword maka term tidak akan diproses lebih lanjut, tapi jika term tidak termasuk dalam daftar stopword maka term akan diproses lebih lanjut. Contoh stopwords adalah “yang”, “dan”, “di”, “dari” dan seterusnya.

Modul Pengubahan Kata Dasar (Stemming). Proses stemming adalah tahap mencari kata dasar (root) dari tiap kata hasil filtering. Pada tahap ini dilakukan proses pengembalian berbagai bentukan kata ke dalam bentuk kata dasar. Proses stemming pada penelitian ini menggunakan Algoritma Porter Stemmer for Bahasa Indonesia [Tala, 2003]. Stemming digunakan untuk untuk mereduksi bentuk term untuk menghindari ketidakcocokan sehingga dapat mengurangi recall.

Modul Indexing (inverted index). Proses Indexing adalah tahap pengindeksan kata dari koleksi teks yang digunakan untuk mempercepat proses pencarian. Seluruh dokumen dalam koleksi disimpan dalam satu file dengan format tertentu sehingga antara dokumen satu dengan dokumen yang lain bisa dibedakan. Setelah kata telah dikembalikan dalam bentuk kata dasar, kemudian disimpan dalam tabel basis data. Proses indexing menghasilkan database index .Metode yang digunakan dalam penelitian ini adalah

inverted index. Modul Vector Space Model (similarity analysis). Hasil indexing selanjutnya dihitung tingkat kemiripannya dengan query menggunakan metode vector space model.

V. IMPLEMENTASI

5.1 Implementasi Vector Space Model

Modul sistem temu kembali informasi terdiri dari : modul pengumpulan dokumen, modul tokenisasi (tokenizing), modul pembuangan stopword (filtering), modul Pengubahan kata dasar (stemming), modul Pengindeksan kata (indexing), dan modul Vector Space Model (term similarity). Modul Pengumpulan Dokumen.

Proses pengumpulan dokumen-dokumen yang dipilih untuk disimpan dalam korpus. Dokumen-dokumen yang dipilih adalah dokumen teks berbahasa Indonesia. Pada penelitian ini digunakan 500 abstraksi skripsi mahasiswa Universitas Indonesia.

Modul Tokenisasi (tokenizing). Tokenizing adalah proses pemotongan string input berdasarkan tiap kata yang menyusunnya. Umumnya setiap kata yang teridentifikasi atau terpisahkan dengan kata yang lain oleh karakter spasi, sehingga proses ini menggunakan spasi pada dokumen untuk melakukan pemisahan kata. Modul Pembuangan stopword (filtering). Tahap filtering adalah proses pembuangan term yang tidak memiliki arti atau tidak relevan. Term yang diperoleh pada tahap tokenisasi dicek dalam suatu daftar stopword, jika term masuk dalam daftar stopword maka term tidak akan diproses lebih lanjut, tapi jika term tidak termasuk dalam daftar stopword maka term akan diproses lebih lanjut. Contoh stopwords adalah “yang”, “dan”, “di”, “dari” dan seterusnya.

Modul Pengubahan Kata Dasar (Stemming). Proses stemming adalah tahap mencari kata dasar (root) dari tiap kata hasil filtering. Pada tahap ini dilakukan proses pengembalian berbagai bentukan kata ke dalam bentuk kata dasar. Proses stemming pada penelitian ini menggunakan Algoritma Porter Stemmer for Bahasa Indonesia [Tala, 2003]. Stemming digunakan untuk untuk mereduksi bentuk term untuk menghindari ketidakcocokan sehingga dapat mengurangi recall.

Modul Indexing (inverted index). Proses Indexing adalah tahap pengindeksan kata dari koleksi teks yang digunakan untuk mempercepat proses pencarian. Seluruh dokumen dalam koleksi disimpan dalam satu file dengan format tertentu sehingga antara dokumen satu dengan dokumen yang lain bisa dibedakan. Setelah kata telah dikembalikan dalam bentuk kata dasar, kemudian disimpan dalam tabel basis data. Proses indexing menghasilkan database index .Metode yang digunakan dalam penelitian ini adalah inverted index. Modul Vector Space Model (similarity analysis). Hasil indexing selanjutnya dihitung tingkat kemiripannya dengan query menggunakan metode vector space model.

VI. KESIMPULAN

Berdasarkan implementasi contoh kasus keyword atau kata kunci (query) sistem dengan 3 dokumen yang ada, vector

space model menghasilkan Dokumen 2 (D2) sebagai dokumen yang paling mendekati kemiripannya atau tingkat kedekatannya dengan query diikuti dengan Dokumen 1 (D1) dan Dokumen 3 (D3). Recall yang dihasilkan rendah dan presisi yang dihasilkan tinggi, artinya dokumen yang diharapkan muncul dengan tingkat akurasi tinggi dapat ditemukan dengan tepat dan dokumen yang dihasilkan sedikit.

DAFTAR PUSTAKA

Lopez, C., Ribeiro., C. (2010) Using Local Precision to Compare Search Engines in Consumer Health Information Retrieval. SIGIR ’10.

Budi, G.S., Gunawan, I., (2006). Algoritma Porter Stemmer for Bahasa Indonesia untuk Pre-Processing Text Mining Berbasis Metode Market Basket Analisys. UK Petra..

Bunyamin., (2008). Aplikasi IR CATA dengan Metode Generalized Vector Space Model, Jurnal Informatika, Vol. 4. No. 1Komputer, No. 6 Thn. 04. (20 Juni-20 Juli 2004)

Smith, (1991). Software Prototyping: Adoption, Practice and Management.

McGraw-Hill, London. Kurniasi, N. 2014. Konsep Dasar Information Retrieval System / Sistem Temu Kembali Informasi. Materi Kuliah Departemen Ilmu Informasi Dan Perpustakaan Universitas Padjajaran.