

SISTEM REKOMENDASI NILAI MATA KULIAH MENGGUNAKAN METODE *CONTENT-BASED FILTERING*

Puspaningtyas Sanjoyo Adi¹⁾

¹⁾ Teknik Informatika, Universitas Sanata Dharma, Yogyakarta
Kampus III Paingan, Maguwoharjo, Sleman, Yogyakarta 55284 Telp (0274)-883037
e-mail : puspa@staff.usd.ac.id

Abstrak

Sistem rekomendasi merupakan sistem yang bertujuan memperkirakan informasi yang menarik bagi penggunanya dan juga membantu calon konsumen dalam memutuskan barang apa saja yang akan dibelinya. Pada awal semester, setiap mahasiswa mendaftar untuk menempuh suatu mata kuliah. Saat inilah, mahasiswa membutuhkan suatu rekomendasi mengenai nilai yang akan diperolehnya. Penelitian ini bertujuan membuat perkiraan nilai mata kuliah yang akan ditempuh oleh seorang mahasiswa. Metode yang digunakan adalah metode content-based filtering.

Metode content-based filtering diimplementasikan dengan masukan sistem adalah dokumen silabus mata kuliah. Dokumen-dokumen ini selanjutnya diproses penghapusan stop word, stemming dan pengindeksan. Proses pengindeksan menghasilkan sebuah daftar kata dan frekuensinya pada dokumen tersebut. Setelah proses pengindeksan selesai, sistem akan melakukan penghitungan bobot kata dalam semua dokumen dengan algoritma TF-IDF. Menggunakan bobot kata ini, dokumen dapat dimodelkan dalam vektor yang dikenal dengan istilah vector space model. Berdasarkan model ini, setiap mata kuliah dihitung tingkat kemiripannya satu sama lain. Nilai tingkat kemiripan ini selanjutnya digunakan untuk membangkitkan rekomendasi nilai menggunakan algoritma K Nearest Neighborhood.

Sistem ini telah diujicobakan pada 10 mahasiswa T. Informatika USD angkatan 2006. Sistem membangkitkan rekomendasi untuk nilai-nilai semester 3 dan semester 4. Nilai yang dibangkitkan sebanyak 176 buah dengan tingkat akurasi sebesar 53%. Tingkat akurasi ini sangat jelek karena masih belum tepatnya algoritma pembangkitan rekomendasi dalam hal ini algoritma K Nearest neighborhood. Masih diperlukan penelitian lanjutan untuk menyempurnakan sistem ini.

Keyword : content based filtering, sistem rekomendasi

1. PENDAHULUAN

Sistem rekomendasi merupakan sistem yang bertujuan memperkirakan informasi yang menarik bagi penggunanya dan juga membantu calon konsumen dalam memutuskan barang apa saja yang akan dibelinya. Sistem rekomendasi secara khusus merupakan suatu sistem pemberian saran kepada penggunanya dan bersifat personal, berbeda untuk semua pengguna sistem. Dalam suatu perguruan tinggi, mahasiswa belajar dalam suatu sistem kurikulum yang terdiri atas sekumpulan daftar mata kuliah dimana setelah menempuh mata kuliah, mahasiswa akan mendapatkan nilai. Nilai mata kuliah yang akan diterima didasarkan atas kemampuan mahasiswa. Setiap mata kuliah mempunyai karakteristik yang berbeda satu sama lain. Karakteristik mata kuliah tercermin dari silabus dan satuan acara perkuliahan. Sistem rekomendasi nilai mata kuliah ini merupakan sistem yang bersifat personal bagi seorang mahasiswa.

Pada awal semester, setiap mahasiswa mendaftar untuk menempuh suatu mata kuliah. Saat inilah, mahasiswa membutuhkan suatu rekomendasi mengenai nilai yang akan diperolehnya. Nilai suatu mata kuliah diperoleh berdasar kemampuan mahasiswa yang juga sesuai dengan karakteristik mata kuliah tersebut.

Berdasarkan karakteristik suatu mata kuliah, profil seorang mahasiswa dibentuk yang sesuai dengan karakteristik mata kuliah. Penentuan profil mahasiswa sesuai karakteristik mata kuliah merupakan implementasi dari metode *content-based filtering*. Dalam mengikuti perkuliahan, nilai suatu mata kuliah yang sejenis umumnya akan sama untuk seorang mahasiswa. Prinsip korelasi antar mata kuliah menjadi dasar dari metode *content based filtering*.

Berdasar latar belakang di atas, rumusan masalah penelitian ini adalah bagaimana membuat perkiraan nilai mata kuliah yang akan ditempuh oleh seorang mahasiswa dalam sebuah bentuk aplikasi sistem rekomendasi.

2. TINJAUAN PUSTAKA

Sistem rekomendasi merupakan suatu sistem yang bertujuan memperkirakan informasi yang menarik bagi penggunanya [1] dan membantu calon konsumen dalam memutuskan barang apa yang akan dibelinya [8]. Perkiraan informasi ini bersifat personal yang didasarkan atas profil dari pengguna sistem. Profil pengguna umumnya didasarkan atas penilaian menarik-tidaknyanya suatu informasi yang pernah dibaca oleh pengguna. Sistem rekomendasi sudah banyak diterapkan pada berbagai model bursa elektronik. Salah satu metode yang digunakan adalah metode *Content-based Filtering*.

Secara umum, metode content-based filtering membentuk profil penggunanya berdasarkan atribut pembentuk suatu item [4]. Sebagai contoh untuk suatu item dokumen, atribut pembentuknya adalah kata-kata/term yang terdapat pada dokumen tersebut. Parameter pembentuk profil pengguna ini juga diberi nilai bobot berdasarkan kriteria tertentu, misalnya untuk kasus dokumen adalah frekuensi suatu kata/term dalam dokumen tersebut.

Content-based filtering memberikan suatu rekomendasi berdasarkan hasil analisa kemiripan item yang telah dinilai oleh para penggunanya. Metode ini pada awalnya diterapkan dalam sistem rekomendasi dokumen. Variasi metode ditawarkan berdasarkan suatu analisis isi dokumen dan pencarian keteraturan dalam berbagai dokumen. Pendekatan metode lain dilakukan menggunakan metode klasifikasi dengan tujuan memperkirakan golongan suatu item atau teks, misalnya golongan disukai atau tidak disukai. Metode lain juga melakukan klasifikasi dengan melakukan pembobotan suatu item atau teks dalam suatu nilai numerik.

Secara umum, metode content-based filtering mempunyai 2 teknik umum dalam membuat rekomendasi yaitu heuristic-based dan model-based [1]. Cosine similarity, Boolean query, teknik TF-IDF (term frequency-invers document frequency) dan Clustering termasuk dalam golongan heuristic-based sedangkan yang masuk dalam golongan model-based adalah teknik Bayesian classifier & Clustering, Decision Tree dan Artificial Neural Network. Penelitian ini akan menggunakan algoritma TF-IDF (term frequency-invers document frequency) sebagai pembentuk profil pengguna.

Algoritma metode content-based filtering dapat dijelaskan dalam langkah-langkah :

1. Suatu item barang dibagi-bagi berdasarkan suatu vektor komponen pembentuknya. Misalnya untuk sebuah film dibagi atas komponen aktor, sutradara, jenis film, dll.
2. Pengguna memberi penilaian suka atau tidak suka atas item tersebut.
3. Sistem akan membuat profil pengguna berdasarkan bobot vektor komponen pembentuk suatu item. Pembuatan profil pengguna dapat menggunakan algoritma TF-IDF (term frequency-invers document frequency). TF adalah jumlah term dalam suatu dokumen. Sedangkan nilai IDF dapat dihitung menggunakan rumus:

$$idf_i = \log\left(\frac{n}{df_i}\right) \dots\dots\dots(1)$$

n merupakan jumlah semua dokumen sedangkan df adalah jumlah dokumen yang memiliki term i .

4. Berdasarkan profil pengguna tersebut, sistem akan memperkirakan penilaian suka atau tidak suka suatu item berdasarkan analisis kemiripan profil pengguna dengan vektor komponen pembentuk item. Jika sistem memperkirakan bahwa item tersebut akan disukai oleh pengguna maka item tersebut akan direkomendasikan ke pengguna.

Metode ini memiliki kekurangan yang utama yaitu ketidakmampuan merekomendasikan jenis item yang baru atau belum pernah dilihat kepada seorang pengguna. Hal ini disebabkan oleh karena metode ini dibuat berdasarkan item-item yang pernah dinilai oleh pengguna tersebut.

Tujuan penelitian inilah adalah mengembangkan suatu perangkat lunak yang dapat membuat perkiraan nilai dari mata kuliah yang ditempuh oleh seorang mahasiswa. Masukan sistem adalah daftar nilai terbaik dan daftar mata kuliah yang sedang ditempuh mahasiswa. Masukan lainnya adalah silabus dan SAP(satuan acara perkuliahan). Studi kasus yang digunakan adalah program studi Teknik Informatika Universitas Sanata Dharma. Hasil penelitian diharapkan mampu membantu mahasiswa dalam mengenali kemampuan diri sendiri. Karakteristik mahasiswa diberikan dalam bentuk term/kata penting kuliah yang dikuasainya dengan baik.

3. METODE PENELITIAN

Langkah Kerja

Metode penelitian yang diacu menggunakan metode pengembangan perangkat lunak berbasis obyek yang bersifat iteratif. Langkah-langkah metode pengembangan perangkat lunak secara umum dapat dilihat pada tabel 1.

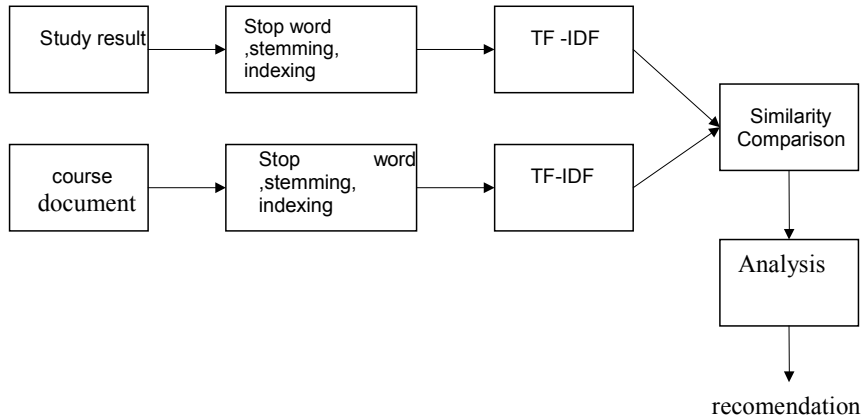
Tabel 1. Langkah-langkah Pengembangan Perangkat Lunak

No	Pekerjaan	Dokumen Keluaran
1.	Analisis kebutuhan	Dokumen Spesifikasi dan Kebutuhan
2.	Analisis dan perancangan sistem	Dokumen Teknis Pengembangan Perangkat Lunak
3.	Pengembangan sistem	Kode Program
4.	Pengujian sistem	Dokumen Pengujian Perangkat Lunak

Proses Pembangkitan Rekomendasi

Perangkat lunak akan dikembangkan berdasarkan proses pembangkitan rekomendasi seperti gambar 1. Masukan utama sistem adalah dokumen SAP/silabus kuliah dan daftar nilai mahasiswa. Keluaran sistem adalah perkiraan/rekomendasi nilai kuliah. Semua dokumen SAP akan diolah dengan cara menghilangkan kata-kata

yang tidak penting (stop word), lalu diambil kata dasarnya (proses stemming). Proses indexing merupakan proses penghitungan jumlah term dalam suatu dokumen sedangkan proses TF-IDF adalah penghitungan bobot dari setiap term pada suatu dokumen. Bobot inilah yang akan digunakan sebagai dasar pembentukan model vektor dari dokumen.

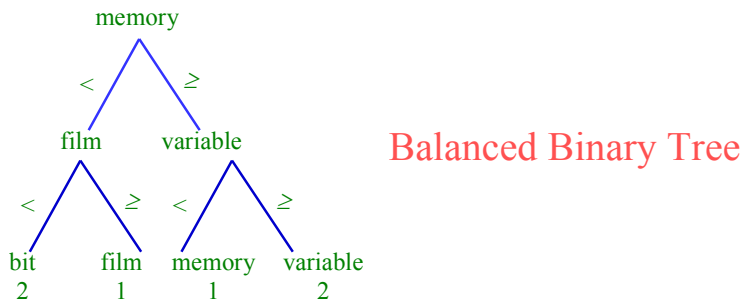


Gambar 1. Proses pembangkitan rekomendasi

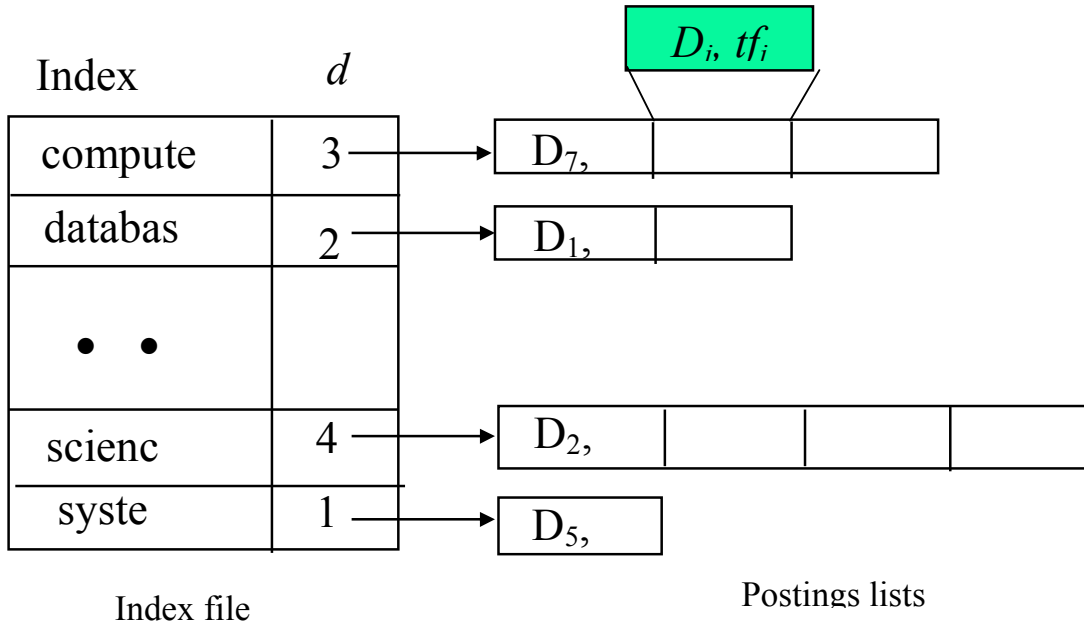
Proses Penghapusan Stop Word, Stemming dan Indexing

Proses ini merupakan proses untuk menghilangkan kata-kata yang tidak penting seperti yang, adalah, dan, dll. Proses stemming merupakan proses untuk menemukan kata dasar. Misalnya, kata 'menghitung' mempunyai kata dasar 'hitung'. Algoritma stemming yang digunakan adalah algoritma Nazief & Adriani's. Pada penelitian ini, satu mata kuliah diwakili oleh satu buah dokumen silabus. Dokumen silabus memuat identitas kuliah, tujuan instruksional umum, pokok bahasan dan sumber pustaka. Setelah sebuah dokumen selesai distemming, proses selanjutnya adalah pembuatan indeks untuk dokumen tersebut. Pada penelitian ini, metode yang digunakan adalah gabungan antara penggunaan mesin basis data MySQL dan pohon biner seperti gambar 2. Pohon biner dipilih karena kecepatan dan kemudahan implementasinya.

Database MySQL digunakan untuk menyimpan kamus kata-kata bahasa Indonesia dan stop word. Database MySQL juga digunakan untuk menyimpan seluruh indeks dari keseluruhan dokumen silabus. Daftar mata kuliah beserta dokumen silabus dapat dilihat pada bagian hasil dan pembahasan. Indeks sebuah dokumen akan berisi informasi kata-kata dan jumlah kata yang terdapat dalam sebuah dokumen. Data indeks ini akan digunakan sebagai vektor pembentuk query masukan. Dengan menggabungkan seluruh indeks dokumen, sistem rekomendasi akan membentuk suatu inverted index (lihat gambar 3). Model inverted indeks ini nantinya akan digunakan untuk menghitung IDF (*invers document frequency*).



Gambar 2. Pohon biner untuk pembentuk indeks dokumen



Gambar 3. Model inverted indeks

Proses Penghitungan Bobot Term dengan TF-IDF

Penghitungan bobot term/kata pada dokumen menggunakan algoritma TF-IDF. Variabel TF merupakan jumlah suatu term/kata dalam suatu dokumen, sedangkan IDF merupakan invers document frequency dari sebuah term/kata yang dapat dihitung menggunakan rumus 1. Sedangkan bobot kata pada suatu dokumen dihitung berdasarkan rumus 2.

$$w_{i,d} = tf_{i,d} \times \log\left(\frac{n}{df_i}\right) \dots\dots\dots(2)$$

$tf_{i,d}$ adalah jumlah term dalam sebuah dokumen.

Dengan menggunakan bobot TF-IDF, sebuah dokumen dapat dimodelkan sebagai sebuah vektor (lihat gambar 4). Dokumen D_i dapat dimodelkan atas komponen T_i sehingga jika seluruh dokumen dikumpulkan maka akan terbentuk matriks term-dokumen dengan nilai bobot term/TF-IDF sebagai nilainya.

$$\begin{pmatrix} & T_1 & T_2 & \dots & T_t \\ D_1 & w_{11} & w_{21} & \dots & w_{t1} \\ D_2 & w_{12} & w_{22} & \dots & w_{t2} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ D_n & w_{1n} & w_{2n} & \dots & w_{tn} \end{pmatrix}$$

Gambar 4. Model vektor dokumen

Proses Penghitungan Kemiripan (Similarity)

Penghitungan kemiripan antar dokumen dilakukan dengan cara menghitung Cosine Similarity antara vektor dokumen koleksi dan vektor dokumen query (lihat gambar 5). Sebuah dokumen dan query akan dimodelkan dalam vektor D dan Q berdasarkan atas nilai bobot semua term (lihat gambar 4). Tingkat kemiripan vektor D dan Q diukur berdasarkan perhitungan Cosine Similarity sbb:

$$\text{CosSim}(d_i, q) = \frac{\vec{d}_i \cdot \vec{q}}{|\vec{d}_i| \cdot |\vec{q}|} = \frac{\sum_{i=1}^t (w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^t w_{ij}^2 \cdot \sum_{i=1}^t w_{iq}^2}} \dots\dots\dots(5)$$