IMPLEMENTASI NAÏVE BAYES CLASSIFIER PADA PROGRAM BANTU PENENTUAN BUKU REFERENSI MATAKULIAH

Atri Nurani[1], Budi Susanto[2], Umi Proboyekti[3]

Abstrak:

Perpustakaan adalah bagian yang penting dari suatu Universitas karena menyediakan buku-buku referensi. Kesulitan yang terjadi adalah ketika perpustakaan harus mengidentifikasi buku-buku referensi tersebut sesuai dengan matakuliahnya. Ada beberapa buku yang sering dijadikan referensi bersama atas beberapa matakuliah. Ada juga buku-buku yang dijadikan referensi tunggal suatu matakuliah, tetapi bahasan materi matakuliah yang bersangkutan tidak dibahas secara optimal dalam buku referensi tersebut. Setiap matakuliah memiliki silabus perkuliahan yang berisi materi-materi dan disusun berdasarkan buku-buku referensi utama dan referensi pendukung dari matakuliah tersebut. Proses klasifikasi akan dilakukan menggunakan metode *Naive Bayesian Classifier (NBC)*. Dalam penelitian ini, proses klasifikasi buku referensi buku menggunakan metode Naive Bayesian memiliki nilai presisi 63%. Dalam melaksanakan tugasnya untuk mengklasifikasikan daftar isi buku referensi sistem dipengaruhi oleh berbagai faktor seperti pola data dan jumlah data training.

Kata Kunci: *Naive Bayesian Classifier*

1. Pendahuluan

Setiap matakuliah memiliki silabus perkuliahan yang berisi materi-materi mengenai matakuliah tersebut. Silabus disusun berdasarkan buku-buku referensi utama dan referensi pendukung dari matakuliah tersebut. Perpustakaan adalah bagian yang penting dari suatu Universitas karena menyediakan buku-buku referensi untuk tiap matakuliah. Kesulitan yang terjadi adalah ketika perpustakaan harus mengidentifikasi buku-buku referensi tersebut sesuai dengan matakuliahnya.

Ada beberapa buku yang sering dijadikan referensi bersama atas beberapa matakuliah. Ada juga buku-buku yang dijadikan referensi tunggal suatu matakuliah, tetapi bahasan materi matakuliah yang bersangkutan tidak dibahas secara optimal dalam buku referensi tersebut. Seringkali judul suatu buku dijadikan gambaran umum mengenai isi suatu buku, padahal isi dari buku tersebut dapat jadi menjelaskan hal yang lain. Daftar isi buku merupakan gambaran khusus dari isi suatu buku. Dari melihat daftar isi, dapat diketahui materi-materi apa saja yang dibahas dalam buku tersebut.

Pada penelitian ini akan dilakukan pengklasifikasian buku-buku referensi berdasarkan silabus matakuliah dengan memanfaatkan informasi dari buku berupa daftar isi. Proses klasifikasi akan dilakukan menggunakan metode Naive Bayesian Classifier (NBC).

Dalam mengkategorikan buku-buku referensi sebagai pendukung matakuliah tertentu atau beberapa matakuliah tertentu maka penelitian ini berfokus pada beberapa

[1] *Atri Nurani, Mahasiswa Teknik Informatika, Fakultas Teknik, Universitas Kristen Duta Wacana*
[2] *Budi Susanto, S.Kom., M.T., Dosen Teknik Informatika, Fakultas Teknik, Universitas Kristen Duta Wacana*
[3] *Umi Proboyekti, Dosen Teknik Informatika, Fakultas Teknik, Universitas Kristen Duta Wacana*

hal yaitu:

a. Bagaimana melakukan klasifikasi yang berdasarkan silabus matakuliah dengan menggunakan informasi dari buku berupa daftar isi?
b. Bagaimana akurasi klasifikasi yang dilakukan berdasarkan pembobotan vektor yang diperoleh dari cocok tidaknya frase tersebut dengan tabel vektor?
c. Bagaimana melakukan klasifikasi terhadap buku yang dijadikan referensi bersama untuk beberapa matakuliah?
d. Bagaimana akurasi dari metode Naïve Bayesian Classifier dalam melakukan klasifikasi pada kasus penentuan buku referensi menggunakan data berupa daftar isi buku?

2. Rancangan Sistem

Data yang digunakan dalam penelitian ini adalah silabus dari 45 matakuliah. Pemilihan 45 matakuliah ini dilakukan berdasarkan hubungan yang ada antara matakuliah-matakuliah tersebut. Selain itu, matakuliah-matakuliah tersebut merupakan inti dari Program Studi Teknik Informatika. Pemilihan 5 matakuliah wajib dari total matakuliah wajib 35 matakuliah adalah dengan pertimbangan bahwa dari 5 matakuliah tersebut berkaitan erat dengan matakuliah-matakuliah konsentrasi. Dapat dikatakan bahwa matakuliah wajib tersebut merupakan sumber dari matakuliah-matakuliah konsentrasi. Sedangkan untuk matakuliah bebas merupakan pengembangan matakuliah konsentrasi.

Adanya kesinambungan antara matakuliah-matakuliah tersebut menyebabkan ada kemiripan materi yang dibahas didalamnya. Dengan begitu, sangat memungkinkan menemukan frase- frase yang sama di matakuliah yang berbeda. Sebagai contohnya adalah *computer network,* secara spesifik *computer network* dibahas pada matakuliah Jaringan Komputer. Tetapi pada matakuliah *Router* dan *Routing* Dasar, *Bridging* dan *Switching* atau matakuliah lain yang membahas mengenai jaringan juga akan membahas *computer network.*

Dari tiap matakuliah telah dilakukan observasi manual dan diambil 5 frase unik. Pemakaian 5 frase untuk tiap matakuliah dianggap cukup untuk membedakan mana frase yang mewakili materi dan mana frase yang berhubungan dengan matakuliah lain. Untuk selanjutnya, daftar frase ini akan digunakan sebagai daftar frase untuk dasar pembobotan data training dan data test.

3. Tinjauan Pustaka

a. *Data Mining*

Data mining mempunyai pengertian sebagai proses penemuan pengetahuan yang bermanfaat dan menarik di dalam kumpulan data yang besar (Jiawei Han dan Micheline Kamber, 2001:5). Tujuan utama *data mining*, yaitu prediksi (*prediction*) dan uraian (*description*). Beberapa tugas utama dari *Data mining* antara lain (Mehmed, 2003:2) adalah *classification* (klasifikasi), *regression* (regresi) , *clustering* (pengelompokan), *summarization* (ringkasan) , *dependency modeling* (pemodelan ketergantungan), *change and deviation detection* (pendeteksian perubahan dan deviasi).

b. *Text Mining*

Text mining adalah bidang multi disiplin yang melibatkan information retreival, text analysis, information extraction, clustering, categorization, visualization, machine learning dan teknik lainnya. (Mehmed, 2003:189). Text mining melakukan ekstraksi informasi terhadap data tesktual (natural language) yang tidak terstruktur, contohnya

dokumen. Text mining menggunakan penerapan data mining untuk mengubah data tidak terstruktur menjadi data terstruktur melalui tahap-tahap yaitu :

1. *Text Preprocess* yaitu pemecahan sekumpulan karakter ke dalam kata-kata
2. *Feature Generation / Text Transformation yaitu* mengubah kata-kata ke dalam bentuk dasar sekaligus mengurangi jumlah kata-kata tersebut.
3. *Feature Selection* yaitu seleksi *feature* untuk mengurangi dimensi dari suatu kumpulan teks.
4. *Text Mining/Pattern Discovery* yaitu dapat berupa *unsupervised learning (clustering)* atau *supervised learning (classification).*
5. *Interpretation/Evaluation* yaitu pengukuran efektifitas untuk mengevaluasi metode yang diterapkan menggunakan parameter *precision.*

4. Landasan Teori

a. *Naïve Bayesian Classifier (NBC)*

NBC menggunakan pendekatan probabilitas untuk menghasilkan *classifier. NBC* menggunakan gabungan probabilitas kata/*term* dengan probabilitas kategori untuk menentukan kemungkinan kategori bagi dokumen yang diberikan. Berikut ini adalah penjelasan mengenai *NBC* (Jiawei Han dan Micheline Kamber, 2001:297)

1) Setiap data direpresentasikan sebagai vektor berdimensi-n yaitu $X=(x_1,x_2,x_3.....,x_n)$,n adalah gambaran dari ukuran yang dibuat di test dari *n* atribut yaitu $A_1,A_2,A_3.....,A_n$
2) m adalah kumpulan kategori yaitu $C_1,C_2,C_3.....C_m$. Diberikan data test *X* yang tidak diketahui kategorinya, maka classifier akan memprediksi bahwa *X* adalah milik kategori dengan posterior probability tertinggi berdasarkan kondisi *X*. Oleh karena itu, *NBC* menandai bahwa test *X* yang tidak diketahui tadi ke kategori C_i jika dan hanya jika :

   $P(C_i|X)>P(C_j|X)$ untuk $1 \le j \le m, j \ne i$

   Kemudian kita perlu memaksimalkan $P(C_i|X)$.

   $$P(C_i|X)= \frac{P(X \mid C_i).P(C_i)}{P(X)}$$

3) $P(X)$ adalah konstan untuk semua kategori, hanya $P(X|C_i).P(C_i)$ yang perlu dimaksimalkan. Jika *prior probability* kategori tidak diketahui, maka akan diasumsikan sama dengan hasil dari kategori-kategori yang lain seperti $P(C1)=P(C2)=.....P(C_m)$ dan oleh karena itu kita akan memaksimalkan $P(X|C_i)$. Sebaliknya, kita memaksimalkan $P(X|C_i).P(C_i)$. Catat bahwa kategori prior probabilities mungkin diperkirakan dengan perhitungan $P(C_i)= \frac{s_i}{s}$ dimana s_i adalah jumlah dari data training dari kategori C_i dan *s* adalah adalah jumlah total data training.
4) Diberikan data dengan banyak atribut, ini akan menjadi komputasi yang kompleks untuk mengkomputasi $P(X|C_i)$. Untuk mengurangi komputasi pada saat mengevaluasi $P(X|C_i)$, maka dapat dihitung menggunakan perhitungan :

   $$P(X|C_i)= \prod_{k=1}^{n} P(x_k \mid C_i)$$

   Dimana x adalah nilai-nilai atribut dalam sampel *X* dan probabilitas $P(x_1|C_i)$, $P(x_2|C_i)$,....., $P(x_n|C_i)$ dapat diperkirakan dari data training.

5. Analisis Sistem

Data data test berjumlah 100 dokumen, diperoleh nilai presisi yaitu 63%. Hasil presisi 63% menyatakan ada 63 dokumen dikategorikan benar dan 4 dokumen yang dikategorikan salah. Jumlah dokumen yang tidak dapat dikategorikan adalah 33 dokumen. Nilai presisi dihitung dengan cara dokumen yang diklasifikasi benar/jumlah dokumen test = 63/100 * 100 %= 63%. Jumlah dokumen yang tidak dapat dikategorikan dianggap sebagai dokumen yang salah.

Adanya beberapa dokumen yang tidak dapat dikategorikan, disebabkan karena hasil $Pr(x|class=n)$ dengan n untuk semua kategorinya bernilai 0. Hasil ini diperoleh dari $Pr(w|class)*Pr(class)$ dengan w adalah frase dan *class* adalah kategori. Jadi apabila salah satu saja dari $Pr(w|class)$ ada yang bernilai 0, maka hal inilah yang akan menimbulkan $Pr(x|class)$ akan bernilai 0. Untuk jumlah kategori yang banyak, kemungkinan munculnya $Pr(w|class)=0$ akan semakin besar karena untuk 2 kategori saja hal ini bisa terjadi. Selain itu, vektor yang dihasilkan dari data training juga memiliki kemungkinan untuk $Pr(w|class)=0$.

Berikut merupakan contoh dari buku yang dijadikan referensi bersama untuk matakuliah Sistem Pakar dan Pengantar Kecerdasan Buatan.

a. ISBN : 0672224437
   Judul : Crash Course in Artificial Intelligence and Expert Systems
b. ISBN : 013482928X
   Judul : Introduction to Artificial Intelligence and Expert System

Untuk buku-buku ini, hasil klasifikasinya yaitu tidak dapat dikategorikan. Berikut akan diberikan contoh mengenai buku yang cocok dijadikan referensi beberapa kategori.

a. Buku dengan ISBN 9630573199 berjudul The Behaviour and Simplicity of Finite Moore Automata. Jika dilihat dari judul bukunya, buku ini adalah buku untuk matakuliah Teori Bahasa Otomata. Setelah diproses dengan sistem penentuan buku, maka sistem menyarankan bahwa buku ini cocok dijadikan referensi untuk matakuliah Teori Bahasa dan Otomata, Teknik Kompiler, Pengolahan Bahasa Natural, Eksperimental Robotika, dan Pemrograman Kecerdasan Buatan. Jika dilihat dari nilai probabilitas tertinggi, buku ini lebih cocok digunakan sebagai referensi matakuliah Teori Bahasa dan Otomata.
b. Buku dengan ISBN 1587050552 berjudul Cisco WAN Switching Professional Reference. Jika dilihat dari judul bukunya, buku ini adalah buku untuk matakuliah Bridging dan Switching dan matakuliah Teknologi WAN. Setelah diproses dengan sistem penentuan buku, maka sistem menyarankan bahwa buku ini cocok dijadikan referensi untuk matakuliah Teknologi WAN dan Pemeliharaan Jaringan. Jika dilihat dari nilai probabilitas tertinggi, buku ini lebih cocok digunakan sebagai referensi matakuliah Teknologi WAN.
c. Buku dengan ISBN 0135995728 berjudul Computer Graphics Mathematical First Steps. Jika dilihat dari judul bukunya, buku ini adalah buku untuk matakuliah Grafika Komputer. Setelah diproses dengan sistem penentuan buku, maka sistem menyarankan bahwa buku ini cocok dijadikan referensi untuk matakuliah Grafika Komputer , Pengolahan Citra Digital, dan Multimedia Internet. Jika dilihat dari nilai probabilitas tertinggi, buku ini lebih cocok digunakan sebagai referensi matakuliah Grafika Komputer.
d. Buku dengan ISBN1584882441 berjudul A First Course in Fuzzy and Neural Control. Jika dilihat dari judul bukunya, buku ini adalah buku untuk matakuliah Jaringan Syaraf Tiruan dan Logika Samar. Setelah diproses dengan sistem penentuan buku, maka sistem menyarankan bahwa buku ini cocok dijadikan referensi untuk

matakuliah Jaringan Syaraf Tiruan Pemrograman Kecerdasan Buatan, Logika Samar ,dan Pengantar Kecerdasan Buatan. Jika dilihat dari nilai probabilitas tertinggi, buku ini lebih cocok digunakan sebagai referensi matakuliah Jaringan Syaraf Tiruan.

Dengan begitu, sistem penentuan buku ini dapat digunakan untuk mengkategorikan buku-buku yang jadikan referensi bersama atau buku-buku yang memiliki bahasan materi untuk beberapa kategori. Dalam penelitian ini, telah dibuat sistem penentuan buku yang tidak menggunakan teori probabilitas sebagai dasar klasifikasinya. Sistem tersebut menggunakan perhitungan jumlah frase untuk tiap kategori sebagai dasar klasifikasinya.

6. Kesimpulan

Dari penelitian yang dilakukan menggunakan metode klasifikasi *Naïve Bayesian Classifier* untuk kasus penentuan buku referensi matakuliah maka dapat ditarik kesimpulan sebagai berikut :

a. Klasifikasi menggunakan metode *Naive Bayesian Classifier* untuk program bantu dapat dilakukan pada kasus ini dengan hasil presisi yang diperoleh adalah 63%
b. Metode Bayesian memerlukan pengetahuan awal untuk dapat mengambil suatu keputusan. Tingkat keberhasilan metode ini sangat tergantung pada pengetahuan awal yang diberikan.
c. Untuk buku-buku yang dijadikan referensi bersama dapat mengklasifikasikan dengan baik berdasarkan nilai probabilitas tertingginya.

7. Daftar Pustaka

Chayo, Yosafat. 2005. **Membuat Aplikasi Point of Sales dengan Microsoft Visual Studio.NET 2005.** Jakarta : PT.Elex Media Komputindo.

Han, Jiawei. Kamber ,Micheline. 2001.**Data Mining: Concepts and Technique**. San Fransisco : Morgan Kaufmann Publishers.

Hearst, Marti. 17 Oktober 2003. **What is text mining?.** Http://www.sims.berkeley.edu/~hearst/text-mining.html

Kantardzic Mehmed. 2003.**Data Mining - Concepts, Models, and Algorithms**. New Jersey: Penerbit IEEE.

Mitchell, Tom M.1997. **Machine Learning**. Singapore: McGraw Hill

Susanto, Budi.2006. **Studi Email Mining : Email Clustering**. Institut Teknologi Bandung

Weiss Sholom M., Nitin Indurkhya, Tong Zhang, Fred J. Damerau. 2005. **Text Mining Predictive Methods for Analyzing Unstructured Information**. Springer

Yung, Kok. 2005. **Membangun Aplikasi Database Dengan Visual Basic. NET 2005 dan Perintah SQL**, Jakarta : PT Elex Media Komputindo.