Rekayasa Text Mining Guna Membantu Referensi Pencarian Daftar Pustaka Menggunakan Metode Bayes

Winarno
Program Studi S1 Teknik Informatika
Fakultas Ilmu Komputer
Universitas Dian Nuswantoro Semarang
Email : arnowinn@gmail.com

***Abstract*— Book is a reference source that is widely used and much sought after. To search the book, keywords used to find the points of information. In the process of finding frequent difficulties when using a keyword search, still can not find a reference book because most of the existing book search system uses a particular query in the database. To help facilitate and maximize the required book search system that is able to understand the meaning of search keywords. With the table of contents as a representative of the book content, it will be developed text mining method for word tracking process on similarity of contents and the process of determining the probability of similarity with the Bayes method, whereby the required reference books are expected to be found. So the reference books can be selected, viewed by the probability of the keyword in the table of contents.**

Keywords: Text Mining, Bayes Method, Search Book.

I. PENDAHULUAN

Pada umumnya dalam penulisan karya ilmiah, mahasiswa perguruan tinggi banyak memanfaatkan internet dan perpustakaan untuk mencari ide serta kumpulan pustaka. Namun demikian buku lebih banyak di jadikan referensi karena dianggap lebih jelas asal sumbernya dibandingkan dari situs internet terkecuali yang berasal dari jurnal *online*. Sehingga mahasiswa banyak memanfaatkan perpustakaan untuk mencari referensi daftar pustaka yang berasal dari buku.

Di lingkungan perguruan tinggi kebutuhan mahasiswa mengenai informasi buku pada saat perkuliahan dan penulisan karya ilmiah sangatlah tinggi. Sumber pendukung penulisan banyak didasarkan pada buku dan jurnal *online*. Namun terjadi kendala oleh mahasiswa yaitu kesulitan dalam mencari referensi buku yang tepat untuk dijadikan bahan pustaka. Tersedianya perpustakaan dan mesin pencari katalog buku belum cukup memenuhi kebutuhan informasi mengenai sumber referensi pustaka, karena dalam pencarian pada katalog buku hanya meyajikan informasi yang terbatas berupa judul buku, nama pengarang, nama penerbit, dan kategori buku.

Dengan demikian dibutuhkan sebuah mesin pencari yang dapat membantu mencari referensi buku sebagai bahan pustaka dengan melihat konten yang dimuat oleh buku. Sehingga dapat ditentukan buku apa saja yang sesuai untuk dijadikan referensi daftar pustaka. Dengan optimalisasi text mining dan metode probabilitas bayes yang diterapkan pada mesin pencaian berbasis web diharapkan dapat membantu menentukan buku referensi yang tepat serta dapat diakses secara mudah melalui internet.

II. IDENTIFIKASI MASALAH

Berdasarkan uraian latar belakang di atas maka dapat diambil kesimpulan permasalahan yaitu tidak semua pengunjung perputakaan atau mahasiswa mengetahui judul buku yang akan dicari. Selain itu banyak yang memasukan sembarang kata kuci pada katalog buku saat pencarian buku, tetapi buku yang dimaksud tidak ditemukan. Maka " Bagaimana merancang pencarian buku yang baik ", yaitu pencarian yang dapat mencari informasi konten sebuah buku, sehingga mampu menyajikan informasi berupa rekomendasi referensi buku yang sesuai dengan kata kunci yang dimaksudkan.

Pada penelitian ini akan direkayasa sebuah sistem berbasis web yang dapat melakukan pencarian buku menggunakan metode sistem *text mining* dan menghitung probabilitas kemunculan kata kunci pencarian menggunakan metode bayes.

Agar penulisan tidak keluar dari pokok penelitian maka pembahasan dibatasi pada rekayasa perangkat lunak untuk *text mining* dan metode bayes. Data yang diolah adalah daftar isi buku.

III. PENELITIAN TERKAIT

Pada penelitian terdahulu dari jurnal teknologi informasi *text mining* dimanfaatkan untuk menentukan klasifikasi informasi seperti klasifikasi berita dan klasifikasi buku berdasarkan kategorinya masing-masing. Metode yang digunakan untuk klasifikasi yaitu NBC (*Naive Bayes Classifier*). Dalam pemprosesan data tekstual bahasa indonesia digunakan algoritma *config-stripping stemmer*. Algoritma *config-stripping stemmer* digunakan dalam proses mengubah sebuah kata turunan menjadi kata dasarnya dengan menggunakan aturan-aturan tertentu.

Disimpulkan bahwa aplikasi sudah mampu melakukan proses klasifikasi data berita secara otomatis dan proses klasifikasi semakin akurat jika data latih yang digunakan dalam pembelajaran berjumlah banyak [2].

Dari penelitian lain oleh Arifin et all (2009) metode bayesian diterapkan sebagai algoritma untuk mencari *keyword*

paper. Pendekatan metode yang digunakan untuk algoritma pencarian yaitu HMAP (*Hypothesis Maximum Appopri Probability*). HMAP menyatakan hipotesa yang diambil berdasarkan nilai probabilitas berdasarkan kondisi prior yang diketahui. HMAP inilah yang digunakan dalam metode bayes untuk proses *machine learning* dari data *training* tertentu.

Berdasarkan hasil analisa dan pengkajian di simpulkan inputan berupa abstrak yang belum memiliki kata kunci diproses menggunakan algoritma bayesian kemudain output menghasilkan kata kunci yang sesuai untuk abstrak[1].

IV. METODE PENELITIAN

A. *Pengumpulan data*

Pada tahapan pengumpulan data dipilih objek penelitian yaitu Perpustakaan Universitas Dian Nuswantoro Semarang. Pada objek penelitian diambil sempel data uji berupa daftar isi buku sebagai data primer. Guna menunjang di perlukan data sekunder berupa studi literatur dari buku dan jurnal online.

B. *Pengembangan sistem*

Tujuan akhir yang akan dicapai adalah merancang aplikasi pencarian buku dengan menerapkan metode text mining dan teorema bayes. Data yang akan diolah adalah daftar isi buku sehingga didapat sebuah pengelolahan data tekstual dalam penerapan text mining. Teorema Bayes diterapakan sebagai pemrosesan pencarian, dimana sajian informasi adalah berupa rekomendasi buku yang memiliki probabilitas terbesar perbandingan daftar isi buku berdasarkan kata kunci pencarian.

V. HASIL DAN PEMBAHASAN

A. *Text Mining*

Penambangan data yang digunakan ditargetkan pada daftar isi text mini ini bertujuan mencari teks yang dianggap penting dan menghilangan teks yang tidak di perlukan tahapan yang di jalankan pada sistem adalah sebagai berikuut :

1. Text preprocessing.
   A. *toLowerCase* merubah karakter huruf menjadi huruf kecil .
   B. *Tekonizing* menghilangkan dilimiter kata yaitu titik(.) , koma (,) , Kurung buka “(” , Kurung tutup “)”, *slash* “/”,*Back slash* “\”.

2. Pembuangan kata sambung bahasa Indonesia

   Guna mempercepat proses pembandingan kata pada saat pencairan, maka menambang data tekstul dibatasi dengan menghilang kata sambung. Kata yang sama dengan kata sambung tidak akan diikutkan dalam proses pencarian diharapkan efisiensi waktu pencarian lebih cepat. Berikut daftar kata sambung dalam bahasa indonesia yang digunakan:

No	Kata	No	Kata
1	dan	10	Sebaliknya
2	Dengan	11	Melainkan
3	Serta	12	Hanya
4	Atau	13	Malah
5	Tetapi	14	Jangankan
6	Namun	15	Kecuali
7	Sedangkan	16	lalu
8	Kemudian	17	Yaitu
9	selanjutnya	18	yakni

Tabel 5.1 tabel kata sambung

Kata sambung ini yang akan digunakan sebagai data pembelajaran yang akan disimpan didalam database.

B. *Metode Bayes*

Pada metode bayes memiliki formula dasar yaitu :

$$P(H|E) = \frac{P(H|E) \times P(H)}{P(E)}$$

. Pada impementasi klasifikasi teks sendiri, metode Bayes memiliki pendekatan formula yang disebut rumus naive bayes Clasifier. Berikut adalah penyederhanaan formula yang akan dijadiakan sebagai premrosesan kata kunci :

$$HMAP = \arg\max_{V_j \in V} P(V_j) \prod_{V \in position} p(a_i|V_j)$$

Dari pendekatan HMAP di atas dijelaskan pendekatan sebagai berikut:

V_j = Buku referensi.

$$P(Vj) = \frac{|docs\ j|}{|exemple\ |}$$

$|docs_j|$ = Kategori pencarian.

$|exemple|$ = Jumlah dokumen kategori pencarian.

(V_j) diasumsikan $\frac{1}{1}$ karena dianggap 1 ketegori dari jumlah 1 ketegori pencarian. Lebih mudahnya digambarkan sebagai berikut.

Buku ==> Referensi ==> Kata kunci

Keterangan diatas menyatakan bahwa buku dapat menjadi kategori referensi dengan syarat mengandung kata kunci. Maka dari analogi sederhana diatas dapat diambil pernyataan bahwa kategorinya adalah 1 dan jumlah dokumen setiap kategori adalah 1. Sehingga $P(V_j)$ diasumsikan $\frac{1}{1}$ dan pada kasus ini nilai ini diabaikan.

Selanjutnya perhitungan $p(a_i|V_j)$ pada setiap buku diterapkan perhitungan berikut :

$$p(a_i|V_j) = \frac{nk+1}{n+|kosakata\ |}$$

a_i = Kata kunci yang terwakili per kata.
n_k= Jumlah kemunculan setiap kata pada setiap buku.
n = Jumlah keseluruhan kemunculan kata.
|kosakata|= Jumlah kata pencarian.

Karena dalam kasus ini yang dilakukan adalah pencarian buku maka jika diketahui n = 0, tidak dilakukan perhitungan pada $p(a_i|V_j)$ dikarenakan tidak ditemukan kata yang sesuai.

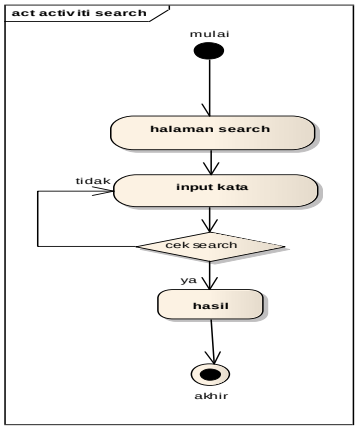
C. *Desain Sistem*

Dijelaskan Desain sistem pokok sebagai berikut :

1. Use Case

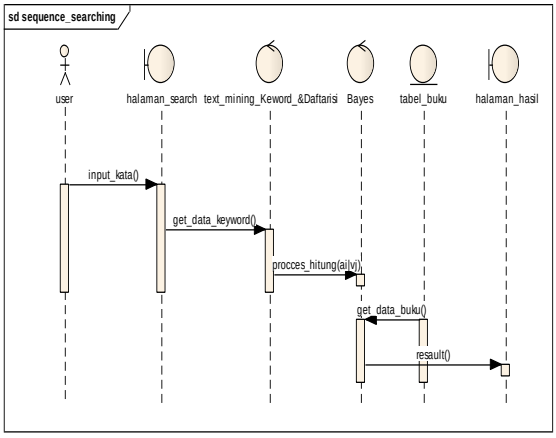
Gambar 5.1 Use Case Diagram

2. Activity Diagram



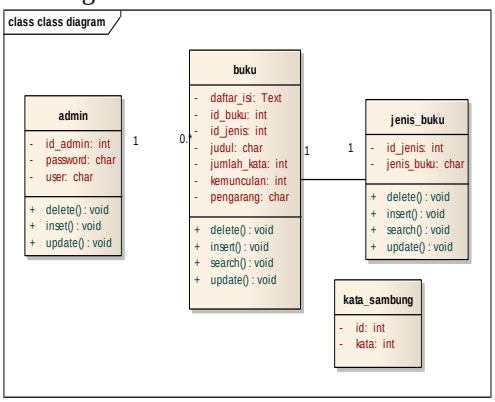
Gambar 5.2 Activity Diagram

3. Sequence Searching



Gambar 5.3 Sequence Diagram

4. Class Diagram



Gambar 5.4 Class Diagram

D. *Hasil implementasi*

Diinputkan judul buku dengan daftar isinya. Input kata kunci pencarian maka akan dicari buku yang memiliki kesamaan kata pada daftar isi sesuai kata kunci selanjutnya akan dihitung probabilitas bayesnya. Dengan metode demikian maka dalam pencarian buku kita dapat mencari buku berdasarkan konten buku tersebut yang diwakili oleh daftar isi buku. Meskipun kata kunci terbalik buku tetap dapat di temukan karena pelacakan berdasarkan kemiripan per kata.

VI. KESIMPULAN

A. *Kesimpulan*

1. Metode *text mining* telah mampu berjalan dan dapat menambang data yang penting sehingga mampu membantu proses pencarian dan perhitungan metode bayes.
2. Dalam metode bayes diperlukan sebuah pengetahuan yang disebut keyword, implementasi yang dilakukan yaitu kata kunci pencarian diasumsikan sebagai pengetahuannya atau keyword.
3. Pemrosesan dilakukan jika kata kunci telah diinputkan dan ditemukan jumlah kata yang sama. Pemrosesna tidak akan dilakukan jika tidak ada pengetahuan atau kata kunci dan jumlah kata yang sama pada kata kunci tidak ada.
4. Hasil pencarin bersifat perkiraan sementara, dan hasil perhitungan sangat bergantung pada kondisi kemunculan setiap kata pada *keyword*.

B. Saran

1. Metode *text mining* masih perlu dikembangkan sehingga penambangan data lebih akurat seperti menambahkan algoritma stemming.
2. Dalam ditambahkannya pengujian *white box* untuk mengetahui akurasi dan kecepatan program.

DAFTAR PUSTAKA

[1] Arifin, F; Hariyadi, M; & Basuki, A. (2009). Pencarian Keyword Paper Menggunakan Algoritma Bayesian. *SemanasIF* , A10-A14.

[2] Kurniwan,B; Efendi,S;& Sitompul, O, S. (2012). Klasifikasi Konten Berita Dengan Metode Text Mining. Jurnal Teknologi Informasi, 14-19.