PENCARIAN KEYWORD PAPER MENGGUNAKAN ALGORITMA BAYESIAN

Firman Arifin[1)], Moch Hariadi[2)], Achmad Basuki[3)]
[1)]Jurusan Elektronika, Politeknik Elektronika Negeri Surabaya
[2)]Jurusan Teknik Elektro,
[3)]Jurusan Teknologi Informasi, Politeknik Elektronika Negeri Surabaya
Institut Teknologi Sepuluh Nopember (ITS) Surabaya
Kampus ITS, Keputih Sukolilo Surabaya
Telepon +62 -31-5947280 Fax +62 -31-6946114
Email : firmanits@yahoo.com

Abstrak
Penggunaan Keyword dalam sebuah informasi sangat diperlukan, karena hal tersebut memudahkan para pembaca untuk mengetahui langsung point-point tertentu yang terdapat pada isi informasi. Karena point-point itu dapat menjelaskan sebagian besar isi informasi dan menyampaikan info kepada pembaca sehingga pembaca tahu apa hal-hal yang bisa ia ambil dari informasi. Pada paper ini menjelaskan bahwa keyword untuk paper atau makalah digunakan untuk mempermudah pembaca mencari point-point penting pada paper atau makalah. Software aplikasi yang kami gunakan adalah Apache (sebagai Web Server), PHP (Pemrograman berbasis Web) dan Mysql (aplikasi Database). Untuk algoritma pemrosesan kami gunakan Bayesian Classifier sebagai machine learning untuk mefilter dan menghasilkan keyword yang sesuai. Program aplikasi pencarian keyword makalah ini akan ditampilkan pada Web, sehingga pengakses mudah untuk menggunakannya

Kata kunci: Keyword, Algoritma Bayesian, Informasi, paper.

1. PENDAHULUAN

Pesatnya perkembangan teknologi informasi pada lingkungan kita, disebabkan karena adanya pemikiran bagaimana aktifitas yang bersifat konvensional dapat dilakukan pada masa sekarang ini dengan proses yang lebih mudah. Dengan muncul teknologi untuk dapat menghubungkan komunitas di seluruh dunia yaitu menggunakan internet, dan layanan ini bisa langsung mengakses sesuai alamat yang akan dituju dengan menggunakan IP Public.

Layanan akses informasi juga cepat dengan adanya bantuan situs pencari seperti Google, Lycos atau Yahoo service, sehingga orang-orang dapat langsung menemukan informasi yang diinginkan tanpa harus membaca semuanya cukup memberikan kata kunci. Disitulah pentingnya kata kunci untuk membantu pencarian informasi yang dibutuhkan. Kata kunci itu juga dapat mewakili point-point penting pada media informasi. Disini kami akan membantu pengakses informasi menemukan kata kunci pada media informasi yang berupa makalah dimaksudkan agar pengakses dapat mengetahui isi pokok dengan melihat kata kunci yang dihasilkan, tanpa harus membaca seluruh isi makalah. Dan pencarian ini menggunakan metode Bayesian karena terbukti metode ini pada machine learning mampu memberikan hasil yang optimal. Dan software yang kami buat berbasis Web dimaksudkan agar semua dapat mengakses dengan mudah karena lewat Internet.

2. TINJAUAN PUSTAKA

Untuk membangun sistem ini maka diperlukan beberapa komponen-komponen sebagai berikut:

2.1 HTML

HTML adalah singkatan dari *Hyper Text Markup Language* ,merupakan bahasa teks yang menggunakan tanda-tanda (*markup*) yang dikenal dengan <tag>, dimana merupakan pengembangan dari SGML (*Standard Generalize Markup Language*). Dengan HTML maka dapatlah dibuat suatu halaman web statis dimana nantinya merupakan dasar dari pembuatan halaman web dinamis. Adapun kelebihannya html ini meliput semua platform, jadi walau anda menggunakan OS *(operating System)* apapun maka akan tetap bisa dijalankan.

2.2 Apache Sebagai Web Server

Web Server (World Wide Web Server) adalah server internet yang melayani koneksi transfer data dalam protocol HTTP (Hypertext Transfer Protocol). Web server saat ini merupakan inti dari server-server internet selain e-mail server, ftp, dan news server. Hal ini dapat dimaklumi karena web server yang telah dirancang untuk dapat melayani berbagai jenis data, mulai dari text, hypertext, gambar (image), suara, plug-in dan sebagainya

Salah satu jenis web server yang banyak dipakai dan digemari adalah Apache. Karena beberapa alasan kemudahan,seperti:

- Bersifat Free (gratis).
- Mudah dalam proses instalasi

- Ringan dalam proses kerja sebagai srver dan cepat dalam proses transfer file.
- Handal dengan berbagai fitur keamanan dan lainnya.
- Bersifat Multiplatform (Perbedaan Operating system masih bisa berkomunikasi).

2.3 PHP pemrograman Web Dinamis

PHP (Hypertext Preprocecor) dikenal sebagai sebuah bahasa skrip yang menyatu dengan tag-tag HTML, diproses hanya di server. Sedangkan hasil yang dikirim ke klien berupa skrip HTML, ditangkap menggunakan browser pada sisi klien. PHP digunakan untuk pembuatan Web Dinamis seperti halnya Active Server Pages (ASP), Java Server Pages (JSP), PERL dan sebagainya. Kelebihan PHP sehingga banyak digemari adalah sebagai berikut:

- Life Cycle yang singkat, sehingga PHP selalu mengikuti perkembangan teknologi Internet.
- Cross platform, php dapat dipakai di hampir semua web server yang ada di pasaran (Apache, AOLServer, fhttpd, Microsoft IIS dan lain-lain), dan dapat dijalankan di berbagai sistem operasi (Windows, Linux, FreeBSD, Unix, Solaris dan sebagainya).
- PHP mendukung banyak paket database baik yang komersil maupun nonkomersil seperti PostgreeSQL, MySQL, Oracle dan lain-lainnya.
- Akses database yang lebih fleksibel.
- Waktu eksekusi lebih cepat.
- Tingkat keamanan tinggi.

2.4 MySQL database Sistem

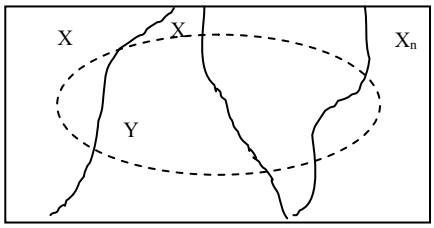
MySQL adalah salah satu jenis database server yang sangat terkenal. Kepopuleran disebabkan karena MySQL menggunakan SQL sebagai bahasa dasar untuk mengakses databasenya. MySQL merupakan server basis data yang menggunakan teknik relasional untuk menghubungkan antara table-tabel dalam databasenya atau mendukung RDBMS (Relational Database Management System), adapun kelebihan lain dari MySQL adalah sebagai berikut:

- MySQL bersifat Free (gratis)
- Dikeluarkan oleh GNU General Public Licence (GPL) sama seperti PHP sehingga keduanya cocok untuk digabungkan.
- MySQL juga bersifat OpenSource jadi para user dapat mengembanggkan pengetahuaanya mengenai MySQL secara gratis.
- Kemampuan yang handal (robust).
- Multi-user (banyak pemakai) cocok untuk server.
- Multi-thread (beberapa prosedur dalam proses dikerjakan bersama) sehingga proses cepat.
- Kecepatan koneksi yang tinggi dan keamanan yang kuat.

2.5 Bayesian Sebagai pemroses Kata Kunci

2.5.1 Teori Bayes

Teori Bayes sebenarnya merupakan implementasi teori probabilitas bersyarat. Teori Bayes seperti probabilitas bersyarat digunakan untuk menentukan probabilitas suatu kejadian Y, bila diketahui kejadian-kejadian lain X_1, X_2, X_3, ..., X_n. Gambaran teori bayes dalam diagram Venn adalah sebagai berikut.



Gambar 1. Diagram Ven Teori Bayes

Probabilitas X_k bila Y diketahui dapat dihitung menggunakan Teori Bayes yang didefinisikan dengan:

$$P(X_k|Y) = \frac{P(Y|X_k)}{\Sigma P(Y|X_i)P(X_i)}$$

2.5.2 HMAP

HMAP (Hypothesis Maximum Appropri Probability) menyatakan hipotesa yang diambil berdasarkan nilai probabilitas berdasarkan kondisi prior yang diketahui. HMAP inilah yang digunakan di dalam metode Bayes untuk proses machine learning dari data training tertentu.

Untuk menentukan HMAP untuk kejadian ya dan tidak dari X, terlebih dahulu diketahui P(X) dan P(~X) yang menyatakan probabilitas X dan probabilitas bukan X. Kemudian diketahui $P(Y_i|X)$ dan $P(Y_i|\sim X)$ yang menyatakan probabilitas Y_i di dalam X dan probabilitas Y_i di dalam bukan X. Data-data itulah yang dinamakan dengan fakta atau dikenal dengan keadaan prior. Dari keadaan prior inilah dapat ditentukan hipotesa yang digunakan untuk menentukan keputusan apakah X atau bukan X. HMAP untuk kejadian S={Y} didefinisikan dengan:

$$P(S \mid X) = \underset{x \in X}{argmax} \frac{P(Y \mid X)\, P(X)}{P(X)}$$

$$= \underset{x \in X}{argmax}\; P(Y \mid X)\, P(X)$$

2.5.3 Implementasi pada klasifikasi Teks

Dengan menggunakan rumus-rumus tersebut maka akan ditentukan algoritma bayesian sebagai pemroses kata kunci, adapun rumus-rumus yang dipakai adalah sebagai berikut:

Rumus Naïve Bayes Clasifier untuk klasifikasi Text, adalah:

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

a_i =Komunitas kata (dimana diwakili (w) untuk perkata)

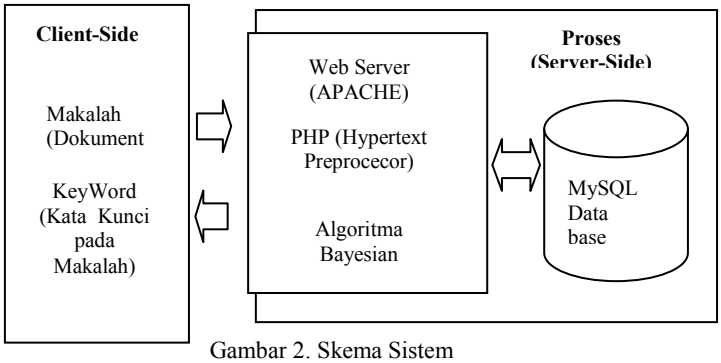
v_j = Kata sebagai target

$$P(v_j) = \frac{|docs_j|}{|example|}$$

$|docs_j|$ = sebagian kata dari kata yang menjadi target Vj

$|example|$ = Jumlah semua kata-kata yang menjadi target Vj.

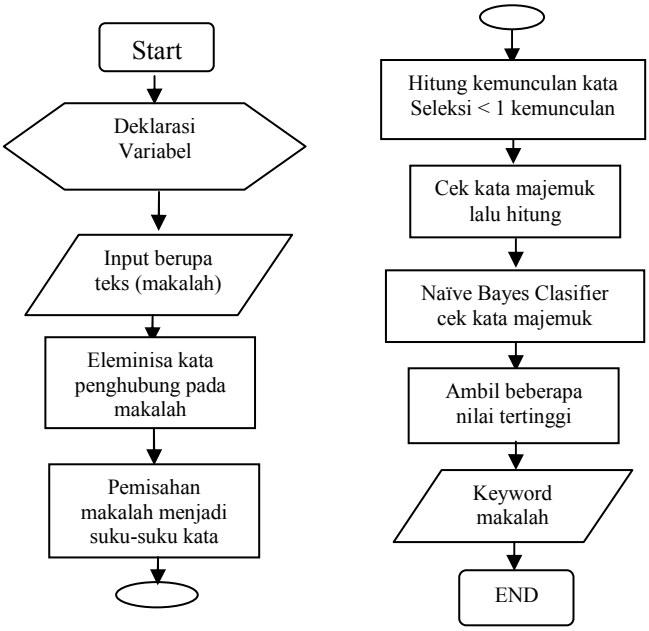
3. METODE PENELITIAN



Gambar 2. Skema Sistem

Dari gambar diatas dapat diketahui bahwa user menginputkan makalah berupa dokumen dengan format teks, masukkan pada halaman web aplikasi sisi klien. Kemudian teks akan dikirim ke sisi server untuk diproses dengan Algoritma Bayesian yang terdapat pada bahasa pemrograman PHP, tidak lupa bekerjasama dengan MySQL sebagai database server. Setelah ditemukan kata kunci, kemudian kata kunci tersebut akan dikirim ke sisi klien, untuk memberi jawaban pada user pemberi inputan tadi.

3.1 Flowchart sistem

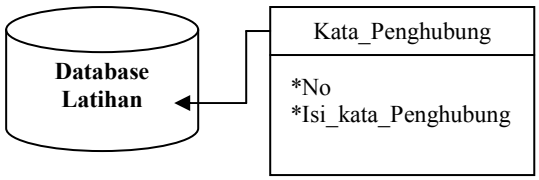


Gambar 3. Flowchart Program

Adapun penjelasan flowchart adalah sebagai berikut:

Di masukkan input berupa teks kedalam variable input, kemudian teks dalam jumlah besar pada makalah itu dipisahkan dengan kategori Kata Penghubung dan Kata Pokok. Jika termasuk salah satu pada salah satu kategori maka akan dimasukkan pada tabel, dimana kata hubung dimasukkan ke tabel kata hubung dan kata pokok dimasukkan ke tabel kata pokok. Tabel kata pokok bersifat sementara hanya untuk menampung kata pokok sebelum diproses menggunakan Algoritma Bayesian. Kemudian kata-kata tersebut dengan metode Naïve Bayes Classifier dipisahkan dengan cara diberi nilai antara True dan false. Jika hasil menunjukkan false data tidak ditampilkan, tapi jika data bernilai true akan ditampilkan, dan kata-kata yang ditampilkan adalah hasil dari proses berupa keyword pada makalah

3.2 Skema Database Sistem



Gambar 4 Skema Database Sistem

Database sistem hanya digunakan untuk meletakkan data kata penghubung, dimana data tersebut digunakan untuk mengeliminasi kata-kata yang tidak perlu sebelum diproses, sebelum algoritma bayesian bekerja.

4.1 HASIL DAN PEMBAHASAN

4.2 Konversi Makalah mejadi foermat Teks

Untuk sementara konversi masih bersifat manual atau tradisional, yaitu dengan:

1) Buka makalah apapun dengan macam-macam format seperti Document, PDF dan lain-lain.
2) Blok semuanya copy atau copy teks jika ada.
3) Buka editor Note Pad, paste dan simpan makalah tersebut.
4) Makalah yang telah disimpan siap untuk diproses.

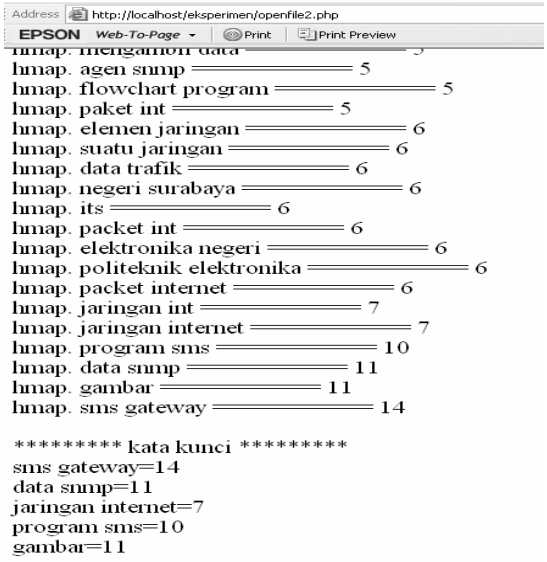
4.3 Buka halaman Web

Buka halaman web tempat aplikasi berada dengan menggunakan Browser pada PC anda, tentunya harus berhubungan dengan server tempat aplikasi berada. Maka akan muncul Form tempat untuk memasukkan input berupa teks makalah. Masukkan makalah dengan menekan tombol Browse, kemudian cari Input makalah dimana anda meletakkannya.

Gambar 5 Tampilan Inputan Pada Browser

4.4 Tampilan Hasil sesudah proses

Kemudian klik tombol proses untuk memproses input dengan algoritma Bayesian



Gambar 6. tampilan Hasil proses beserta Kata Kunci

5. KESIMPULAN

Berdasarkan dari hasil analisa dan pengkajian ini, maka penulis mengambil kesimpulan sebagai berikut:

1. User Menginputkan makalah harus dalam format teks, jika terdapat format lain maka data tersebut tidak akan diproses.
2. Input makalah sebelum diproses dieleminasi kata penghubung beserta karakter selain teks seperti angka, petik dan lain-lain.
3. Sesudah itu akan masuk pada algoritma bayesian untuk memproses makalah kemudian menghasilkan kata kunci makalah tersebut.

6. DAFTAR PUSTAKA

[1] Basuki, Achmad, *"Machine Learning",* PENS-ITS, Surabaya,
[2] Kadir, Abdul*,"Dasar PemrogramanWeb Dinamis Menggunakan PHP",* Penerbit ANDI, Yogyakarta,
[3] M, Farid, "*Belajar Sendiri Pemrograman PHP4*", Elex Media Komputindo, Jakarta, 2001.
[4] Mccallum, Andrew and Nigam, Kamal,*"A Comparison of Event Models for Naïve Bayes Text Classification"*,http://www.cs.cmu.edu
[5] Nur Iman, Budi ,dkk*,"Statistika dan Probabilitas"*, PENS-ITS, Surabaya, 2001.
[6] Shen, Yirong and Jiang, Jing ,*"Improving the performance of Naïve Bayes for Text Classification"*, CS224N Spring,2003.
[7] Wasista ,Sigit, "*Pemrograman Web* ", PENS-ITS, Surabaya, Juni 2002.