

Bab I

Pendahuluan

1.1 Latar Belakang

Perkembangan World Wide Web dalam 10 tahun terakhir ini semakin pesat, dan sekarang keberadaan Internet sudah tidak bisa diabaikan lagi. Dengan banyaknya informasi yang tersedia di Internet, berbagai layanan telah dikembangkan untuk membantu pengguna Internet dalam menemukan informasi yang mereka inginkan.

Layanan yang paling populer adalah mesin pencari atau *search engine*, misalnya Google (<http://www.google.com>), Ask Jeeves (<http://www.ask.com>), dan MSN Search (<http://www.msnsearch.com>), yang dapat digunakan untuk mencari informasi berdasarkan kata-kata kunci tertentu. Meski sangat berguna, kadang-kadang kita ingin mencari informasi berdasarkan kategori tertentu. Untuk itu ada layanan direktori, misalnya Yahoo! (<http://www.yahoo.com>) dan dmoz.org Open Directory (<http://www.dmoz.org>), yang berisi daftar situs Web berdasarkan klasifikasi tertentu. Layanan direktori ini sebenarnya sangat berguna, namun memiliki banyak kelemahan, di antaranya adalah kurang terupdate, struktur kategori yang statis, dan kurangnya staf untuk merawat ribuan kategori dan subkategori yang terdapat dalam sebuah layanan direktori.

Kategori situs yang akhir-akhir ini populer adalah *social networking sites*, misalnya Friendster (<http://www.friendster.com>) dan GaulDong (<http://www.gauldong.net>). Pesatnya perkembangan *social sites* membuat banyak layanan lain bermunculan, salah satunya adalah del.icio.us (<http://del.icio.us>) yang merupakan *social bookmarking site* berdasarkan *folksonomy*. Setiap pengguna del.icio.us bebas mengirim link/bookmark ke sebuah situs yang dikategorikan ke satu atau lebih *tag* (istilah del.icio.us untuk kategori). *Folksonomy* merupakan teknik klasifikasi yang tidak hierarkis, melainkan semua kategori disamaratakan, dan kategori dibuat secara bebas berdasarkan masukan dari pengguna.

Penggunaan *folksonomy*, meski sangat mudah, juga mempunyai beberapa kelemahan, yaitu penggunaan *tag* yang berbeda-beda untuk konsep yang sama, penggunaan *tag* yang sama untuk konsep yang berbeda-beda, tidak adanya pengendalian mutu, dan lain-lain. Di sini penulis mencoba memberikan solusi untuk sebagian masalah tersebut yaitu dengan cara menganalisa isi dari dokumen Web yang ditunjuk dan mengkategorisasikan link tersebut secara otomatis ke beberapa *tag* menggunakan multinomial naive Bayes classifier untuk keperluan ini.

1.2 Perumusan Masalah

Permasalahan yang dijadikan objek penelitian tugas akhir ini adalah menitikberatkan pada analisa penggunaan multinomial naive Bayes classifier untuk kategorisasi dokumen Web secara otomatis. Dari penelitian ini diharapkan dapat diketahui bagaimana performansi algoritma tersebut dalam melakukan kategorisasi otomatis dokumen Web.

1.3 Tujuan

Tujuan atau hasil akhir yang ingin dicapai dari tugas akhir ini adalah:

1. Mengimplementasikan sebuah sistem kategorisasi dokumen Web dengan fasilitas minimal, bernama Gado-gado.
2. Menerapkan multinomial naive Bayes classifier untuk kategorisasi dokumen secara otomatis.
3. Melakukan analisa terhadap akurasi kategorisasi otomatis yang dilakukan dibandingkan dengan kategorisasi secara manual, serta analisa performansi dan efisiensi.

1.4 Batasan Masalah

Untuk menghindari meluasnya materi pembahasan tugas akhir ini, maka penulis membatasi permasalahan dalam tugas akhir ini hanya mencakup hal-hal berikut:

1. Aplikasi akan dibangun dengan menggunakan PHP 5.0.3 sebagai web scripting language, Apache 2.0 sebagai web server, MySQL sebagai database management system, dan Windows XP Professional sebagai operating system.
2. Aktivitas yang dilakukan dalam sistem ini meliputi login user, penambahan link, menampilkan daftar link dalam sebuah tag, dan melakukan uji coba.
3. Dokumen yang dapat diterima hanya dalam bahasa Inggris (variasi apa pun) atau Indonesia (tidak menggunakan dua bahasa dalam satu dokumen), dengan encoding berikut: ISO-8859-1, WIN-1252, atau UTF-8.
4. Ukuran keseluruhan database maksimal 50 MB.

1.5 Metodologi Penyelesaian Masalah

Berikut ini adalah metodologi penyelesaian masalah yang dipergunakan dalam tugas akhir ini:

1. Studi literatur

Bertujuan untuk mempelajari dasar teori dari literatur-literatur tentang :

- Folksonomy
- Indexing
- Inverted index
- Klasifikasi teks
- Teori probabilitas
- Bayes theorem
- Naive Bayes classifier
- Multinomial naive Bayes classifier

2. Pengumpulan data untuk inputan

3. Studi perancangan perangkat lunak

Bertujuan untuk menentukan metodologi pengembangan perangkat lunak yang digunakan dengan menggunakan metode terstruktur dan melakukan perancangan sistem.

4. Pembuatan perangkat lunak

Bertujuan untuk melakukan implementasi metode pada perangkat lunak sesuai dengan analisa perancangan yang telah dilakukan.

5. Pengujian perangkat lunak

6. Analisa terhadap hasil pengujian

7. Pengambilan kesimpulan dan penyusunan laporan

1.6 **Sistematika Penulisan**

BAB I PENDAHULUAN (INTRODUCTION)

Berisi latar belakang, perumusan masalah, tujuan pembahasan., metodologi penyelesaian masalah dan sistematika penulisan.

BAB II LANDASAN TEORI (FUNDAMENTAL THEORIES)

Penjelasan mengenai *folksonomy*, *indexing*, *inverted index*, *text classifier*, teori probabilitas, *Bayes theorem*, *naive Bayes*, *multinomial naive Bayes*.

BAB III ANALISIS DAN PERANCANGAN SISTEM (SYSTEM ANALYSIS AND DESIGN)

Membahas tentang perancangan awal sistem dengan metode Test-Driven Development (TDD) menggunakan bahasa pemodelan UML.

BAB IV IMPLEMENTASI DAN PENGUJIAN (IMPLEMENTATION AND TESTING)

Menyajikan hasil pengujian dan analisa terhadap kategorisasi dokumen Web secara otomatis berdasarkan folksonomy menggunakan multinomial naive Bayes classifier.

BAB V KESIMPULAN DAN SARAN (CONCLUSIONS AND SUGGESTIONS)

Berisi kesimpulan dan saran pengembangan.