BAB 1. Pendahuluan

1.1 Latar Belakang

Halaman web merupakan dokumen yang ditulis dalam format HTML (Hypertext Markup Language) yang biasanya dapat diakses melalui http yaitu sebuah protokol yang menyampaikan dari server website untuk ditampilkan kepada pengguna melalui browser. Klasifikasi halaman web dapat diartikan sebagai bidang penelitian dalam perolehan informasi yang mengembangkan metode untuk menentukan atau mengkategorikan sebuah dokumen yang terdapat dalam suatu halaman web ke dalam satu atau lebih kelompok dokumen.

Perkembangan Internet sudah merambah hampir pada kalangan umum dan sudah menjadi kebutuhan banyak orang karena kita dapat mengakses dan mendapatkan informasi yang cepat dari seluruh dunia. Informasi yang bisa didapatkan melalui internet dapat berupa teks, gambar, video, audio dan komponen multimedia lainnya.

Informasi tersebut dapat diakses melalui halaman web. Web memuat banyak informasi yang dihasilkan dari waktu ke waktu secara kontinu dari berbagai sumber. Jumlah informasi yang terus bertambah dapat menyulitkan para pencari informasi dalam menemukan informasi yang relevan. Salah satu cara yang dirasa efektif untuk menyelesaikan permasalahan ini adalah dengan melakukan klasifikasi halaman web menurut topiknya.

Penelitian sebelumnya telah dilakukan untuk mengklasifikasikan halaman web yang mengelompokan dokumen-dokumen halaman web menurut isinya diantaranya yaitu dengan menggunakan *Naïve Bayes Classifier* dan menggunakan *Ontologi*. Kedua metode ini mempunyai beberapa persamaan yaitu memakai konsep class dan frekuensi term dimana frekuensi ini dihitung menggunakan formula dari masing-masing metode. Dan kedua metode ini hanya menekankan pada isi dokumen saja tetapi tidak memperhatikan informasi penting yang dapat diambil dari dokumen

seperti struktur HTML dan *hypertext link*. Serta umumnya pemakaian kedua metode diatas jika ada penambahan jumlah kategori dapat menyebabkan penurunan tingkat akurasi dalam klasifikasi dokumen pada halaman web.

Sehingga dalam tugas akhir ini diambil sebuah metode yang dapat menyelesaikan permasalahan pada dua metode yang dijelaskan diatas yaitu menerapkan metode *categorization by context*. Metode yang mengkategorisasikan dokumen pada halaman web dengan tidak menggunakan kemampuan menganalisis isi dokumen melainkan mengekstrak informasi yang berguna dari sebuah dokumen halaman web untuk mengklasifikasi dokumen dimana URL muncul sebagai rujukannya[1].

URL muncul sebagai rujukan karena metode ini mengekspoitasi petunjuk di sekitar URL. Petunjuk yang diekploitasi tersebut adalah struktur dokumen HTML dan *hypertext link*. Penambahan jumlah kategori pada metode ini tidak mempengaruhi tingkat keakuratan kategori klasifikasi dokumen pada halaman web. [1].

1.2 Perumusan Masalah

Pada tugas akhir ini terdapat beberapa rumusan masalah agar dapat menghasilkan klasifikasi halaman web menurut topiknya, yaitu :

1. Bagaimana mendapatkan data URL terkait dan struktur dokumen HTML yang diperoleh dari dokumen web HTML.
2. Bagaimana menerapkan metode *categorization by context* dalam menghasilkan klasifikasi halaman web berdasarkan topiknya.

Dalam penelitian tugas akhir ini,obyek penelitian dibatasi dengan ruang lingkup sebagai berikut :

1. Implementasi yang dibuat dalam tugas akhir ini menekankan pada hasil kategorisasi yang didapat dari pengolahan data dokumen pada halaman web dengan memanfaatkan informasi struktur HTML dan URL.

2. Halaman web yang dijadikan data sampel untuk kategorisasi adalah halaman web berbahasa Indonesia dengan mengambil studi kasus pada situs web http://www.okefood.com/,http://www.tempo.co/,http://www.antaranews.com/, http://www.kompas.com/,http://www.republika.co.id/,http://www.bbc.co.uk/indonesia/, http://id.berita.yahoo.com/.
3. Dokumen pada halaman situs-situs web yang dijadikan studi kasus diambil pada rentang waktu tertentu dan disimpan dalam database perangkat lunak.

1.3 Tujuan

Tujuan yang ingin dicapai melalui tugas akhir ini adalah :

1. Dapat mengimplementasikan metode *categorization by context* untuk klasifikasi halaman web
2. Dapat menganalisis performansi dari hasil implementasi yang akan dilakukan pada tugas akhir ini.

1.4 Metodologi Penyelesaian Masalah

Penyelesaian masalah dalam tugas akhir ini adalah :

1. Studi Literatur

   Mempelajari literature-literatur yang berhubungan dengan novel teknik, struktur dokumen HTML, identifikasi link, dan *web page categorization* serta menghimpun informasi yang relevan yang berhubungan dengan klasifikasi web.
2. Pengumpulan Data

   Mengambil sampel data yang akan digunakan pada saat implementasi. Data sampel yang digunakan adalah menggunakan 9 situs berita bahasa indonesia.
3. Perancangan Sistem

   Membuat rancangan sistem dengan pemodelan desain yang sesuai dengan implementasi yang akan dilakukan. Rancangan desain yang akan dilakukan adalah pemodelan terstruktur yaitu desain diagram tahapan proses

menggunakan diagram blok, desain diagram aliran data, kamus data, dan spesifikasi proses.

4. Implementasi

   Melakukan implementasi dengan mengambil data dari data sampel yang telah dikumpulkan sebelumnya dan sesuai dengan model rancangan sistem yang telah dibuat.

5. Analisis hasil

   Melakukan analisis hasil dari hasil implementasi yang telah dilakukan pada tahap sebelumnya.

6. Pembuatan laporan

   Menuliskan detail dari hasil implementasi dan hasil analisis hasil yang dibuat sedemikian rupa sehingga dapat dibaca dan didokumentasikan dengan baik dan benar.

1.5 Sistematika Penulisan

Sistematika penulisan tugas akhir ini adalah sebagai berikut:

1. Pendaluluan

Membahas tentang latar belakang masalah, rumusan masalah, , tujuan penulisan, batasan masalah, metodologi penyelesaian masalah, dan sistematika penulisan.

2. Tinjauan Pustaka

Membahas tentang dasar teori yang berkaitan dengan tugas akhir ini.

3. Perancangan Sistem

Menjelaskan bagaimana tahap-tahap perancangan dan desain sistem yang akan dibangun pada tugas akhir ini.

4. Pengujian dan Analisis Sistem

Melakukan pengujian dan analisis terhadap sistem dalam mengklasifikasikan dokumen halaman web.

5. Kesimpulan dan Saran

Berisi kesimpulan dan saran yang berkaitan dengan tugas akhir ini.