KLASIFIKASI HALAMAN WEB DENGAN ANALISIS KONTEKS DAN URL

Thursina Andini[1], Yanuar Firdaus A.w.[2], Arie Ardiyanti Suryani[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Pencarian informasi dapat diperoleh dari internet dengan mudah dan cepat 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.

Salah satu metode untuk mengklasifikasikan halaman web berdasarkan topiknya yaitu categorization by context. Metode categorization by context mengekstrak informasi yang berguna dari sebuah dokumen halaman web untuk mengklasifikasi dokumen dimana URL muncul sebagai rujukannya. Metode ini dirasa efektif karena tidak mengkategorisasikan dokumen pada halaman web menurut isinya namun berdasarkan URL terkait dan tag yang ada pada suatu dokumen

Sebuah sistem berbentuk sebuah katalog diimplementasikan yang isinya terdiri dari kategori-kategori dari halaman web dimana kategori ditentukan oleh penulis sendiri. Pada implementasi ini dokumen-dokumen diklasifikasikan ke dalam kategori tertentu berdasarkan tag dan URL terkaitnya yang kemudian dicocokan dengan parameter setiap kategori yang didapatkan. Hasil yang diperoleh menunjukkan sebagian besar dokumen terklasifikasi sesuai dengan kategori prediksinya.

Kata Kunci : categorization by context, URL terkait, tag

Abstract

Searching specific information can be get from internet easily and fast through web pages. Web pages contain many information thar result from time to time continously from many resources. A big number of information make difficult user to find relevant information. One of effective method that can solve this problem is make web page classification based on the subject.

One of method than can make web page classification based on the subject is categorization by context. This method categorization by context is extracting the useful information form a web page document to classify a document where URL referring to appears. This method can more effective because doesnt categorize a document on web page based on content or text document but based on surrounding links and tag in a document.

A catalog system implemented which contains categories form web page where a category define by author. In this implementation documents classify into certain category based on surrounding links and tag then match with parameter on each category. The result shows that almost documents classified appropriate with their expected category.

Keywords : categorization by context, surrounding links, tag

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.

BAB 5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan hasil pengukuran dan analisis terhadap sistem yang dilakukan pada saat pengujian, dapat diambil kesimpulan sebagai berikut :

1. Sistem yang dibangun untuk mengimplementasikan metode yang diambil dalam tugas akhir ini dapat mengkategorisasikan sebagian besar dokumen sesuai dengan kategori prediksinya dengan hasil *f-measure* 94%.
2. Hasil pengukuran performansi sistem berhasil dilakukan dengan baik. Pengukuran performansi sistem mencakup hal-hal sebagai berikut :
   a) Hasil *recall, precision,* dan *F-measure* menghasilkan nilai yang tinggi disebabkan karena sistem dapat mengenali dengan baik *URL terkait* dan *tag* <*title*> suatu dokumen kemudian berhasil mencocokkannya dengan parameter kategori yang ada pada sistem.
   b) Jadi dapat disimpulkan bahwa parameter yang diperoleh dapat diaplikasikan pada data testing yang diujikan sehingga akurasi menunjukkan hasil yang tinggi serta pemilihan kata stopword harus lebih teliti dan tepat sehingga stopword tidak seharusnya menyebabkan penurunan tingkat akurasi pada sistem.
   c) Dalam pengujian pengaruh jumlah kategori terbukti, semakin bertambahnya kategori tidak menyebabkan penurunan tingkat kecocokan. Hal ini dikarenakan tingkat kemiripan diantara kategori cukup rendah.
   d) Dari beberapa pengujian yang dilakukan terhadap klasifikasi dokumen menunjukkan bahwa tingkat kecocokan pada suatu file uji dapat naik jika kategorinya ada di dalam sistem dan file uji cocok dengan parameter pada kategori yang bersangkutan.

5.2 Saran

Beberapa saran jika akan dilakukan penelitian yang lebih lanjut mengenai tugas akhir ini adalah sebagai berikut :

1. Studi lebih lanjut mengenai metode *categorization by context* yang memungkinkan adanya perkembangan pada metode ini.
2. Lakukan pengujian terhadap struktur tag HTML selain tag<title> terhadap tingkat kecocokan dokumen dengan suatu kategori.
3. Memperbanyak variasi data pada proses *Training* "parameter" agar parameter yang dihasilkan lebih bervariasi.
4. Mencari metode lain untuk klasifikasi dokumen atau perpaduan metode klasfikasi lain agar akurasi sistem lebih meningkat.

Daftar Pustaka

[1] Attardi, Giuseppe., Antonio Gulli., Fabrizio Sebastiani. 2000. *Automatic Web Page Categorization by Link and Context Analysis*. Italy : Universita di Pisa.

[2] Basnur, WiraPrajna., Dana IndraSensuse. 2010. *Pengklasifikasian Otomastis Berbasis Ontologi Untuk Artikel Berita Berbahasa Indonesia*. Indonesia : Universitas Indonesia.

[3] Cooley, R., B.Mobasher., J.Srivastava. 2000. *Web Mining : Information and Pattern Discovery on the World Wide Web*. USA : University of Minnesota.

[4] Gozali, Ferrianto., Mochamad Fajar Faezal. 2004. *Peranan Web Spider Dalam Internet Search Engine*. Indonesia : Universitas Trisakti.

[5] Kusnawi. 2007. Pengantar Solusi Data Mining. Indonesia : STMIK AMIKOM.

[6] Pan, Fengsia. 2006. *Multi-Dimensional Fragment Classification in Biomedical text*. Canada : Queen's University.

[7] Budi, Indra., Rizal Fathoni Aji. 2006. *Efektifitas Seleksi Fitur Dalam Sistem Temu Kembali Informasi*. Indonesia : Universitas Indonesia.

[8] Attardi, Giuseppe., Sergio di Marco., Davide Salvi. 1998. *Categorization by Context. Italy* : Universita di Pisa.