EKSTRAKSI CONTENT STRUCTURE PADA HALAMAN WEB MENGGUNAKAN METODE VISION BASED PAGE SEGMENTATION

Andri Setiawan[1], Yanuar Firdaus A.w.[2], Arie Ardiyanti Suryani[3]

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

Abstrak

Sebuah halaman web biasanya mengandung berbagai jenis content seperti navigasi, dekorasi, dan bagian-bagian lain yang tidak berhubungan dengan inti informasi dari halaman web tersebut. Di sisi lain, kadang pengguna sebenarnya hanya membutuhkan informasi inti dari halaman tersebut. Dari sinilah muncul kebutuhan akan sistem yang dapat mengekstrak informasi dari suatu halaman web.

User melihat sebuah halaman web melalui web browser dan mendapatkan representasi 2D yang mempunyai banyak visual cues (penanda visual) untuk membantu membedakan bagian bagian yang berbeda dari halaman tersebut. Seorang web designer biasanya mengorganisasi content dari sebuah halaman web agar mudah untuk dibaca/dipahami oleh user. Oleh karena itu, content-content yang berhubungan secara semantik biasanya diletakkan dalam satu kelompok dan halaman web tersebut dibagi menjadi region-region untuk content yang berbeda dengan menggunakan pembeda visual seperti garis, ukuran font, warna, dll. Content-content yang sejenis biasanya akan ditampilkan dengan bentuk visual yang sama atau sejenis pula. Visual cues inilah yang akan dimanfaatkan untuk proses identifikasi dan ekstraksi data. Metode Visual-Based Page Segmentation akan memanfaatkan penanda visual (visual cues) dari halaman web untuk mengekstrak data dari halaman web tersebut.

Tahap analisis dan pengujian memberikan hasil bahwa pattern visual cues yang tepat terbukti dapat dimanfaatkan untuk membuat sistem ekstraksi informasi dari halaman web meskipun masih terdapat noise.

Kata Kunci : halaman web, ekstraksi, visual cues, pattern, noise

Abstract

A web page usually contains various types of content such as navigation, decorations, and other parts that are not associated with the core information from these web pages. On the other hand, sometimes the user actually requires only core information from these pages. From this came the need for a system that can extract information from a web page.

Users see a web page through a web browser and get a 2D representation that have a lot of visual cues to help distinguish different parts of the page. Web designers usually organize content from a web page so that it is easy to understood by the user. Therefore, the content-related content semantically usually placed in one group and web pages are divided into regions for different content using a visual differentiator such as line, font size, color, etc. Same type content would normally be displayed with a similar visual form as well. These visual cues will be used for identification and data extraction processes.Visual Based Page Segmentation Method will use visual cues from the web page to extract data from these web pages.

Phase analysis and test results provide proves that appropriate pattern of visual cues can be used to create a system of information extraction from web pages although there is still some noises.

Keywords : web page, extraction, visual cues, pattern, noise

1. Pendahuluan

1.1 Latar belakang

Saat ini perkembangan internet sudah sedemikian cepat. Lebih dari 2 milyar halaman web telah di-*publish* sejak tahun 1995, dan bertambah sekitar 200 juta halaman baru setiap bulannya[11]. Perkembangan yang sedemikian cepat ini menggambarkan betapa banyaknya informasi yang tersedia untuk dimanfaatkan oleh pengguna internet. Namun hal ini juga memberikan satu tantangan baru yaitu bagaimana pengguna internet dapat memilih dan menyerap informasi-informasi yang dibutuhkan dengan cepat dan tepat diantara sedemikian banyak atau bahkan dari sebuah halaman web.

Sebuah web page biasanya mengandung berbagai jenis content seperti navigasi, dekorasi, dan bagian-bagian lain yang tidak berhubungan dengan inti informasi dari halaman web tersebut. Di sisi lain, kadang pengguna sebenarnya hanya membutuhkan informasi inti dari halaman tersebut. Dari sinilah muncul kebutuhan akan sistem yang dapat mengekstrak informasi dari suatu halaman web.

Banyak aplikasi web yang dapat memanfaatkan data yang telah diekstrak dari halaman web. Data yang telah diekstrak dari suatu web site dapat diintegerasikan dalam satu *data collection* untuk kemudian diolah sesuai dengan kebutuhan, misalnya untuk sistem pembanding harga produk antar beberapa web toko online. Contoh yang lain, deteksi *content structure* dari sebuah halaman web dapat dimanfaatkan untuk membuat sistem browser untuk *handled device* (sebagai fasilitas untuk *browsing* halaman yang cukup panjang dengan membaginya ke bagian-bagian yang lebih kecil). Selain itu, ada juga yang menyimpan data yang telah diekstrak tersebut ke dalam database untuk kemudian diteliti dengan menggunakan proses-proses data mining untuk keperluan tertentu.

Tugas Akhir ini akan menggunakan metode *Visual-Based Page Segmentation* untuk mengekstrak data dari halaman web. Metode ini akan memanfaatkan penanda visual (*visual cues*) dari halaman web untuk mengekstrak data dari halaman web tersebut.

User melihat sebuah halaman web melalui web browser dan mendapatkan representasi 2D yang mempunyai banyak *visual cues* (penanda visual) untuk membantu membedakan bagian bagian yang berbeda dari halaman tersebut. Seorang *web designer* biasanya mengorganisasi *content* dari sebuah halaman web agar mudah untuk dibaca/dipahami oleh user. Oleh karena itu, *content-content* yang berhubungan secara semantik biasanya diletakkan dalam satu kelompok dan halaman web tersebut dibagi menjadi region-region untuk *content* yang berbeda dengan menggunakan visual separator seperti garis, daerah kosong, gambar, ukuran *font*, warna, dll. Content-content yang sejenis biasanya akan ditampilkan dengan bentuk visual yang sama atau sejenis pula. *Visual cues* inilah yang akan dimanfaatkan untuk proses identifikasi dan ekstraksi data.

1.2 Perumusan masalah

Beberapa permasalahan yang akan dibahas dalam tugas akhir ini diantaranya:

1. Bagaimana mengimplementasikan metode *Vision Based Page Segmentation* untuk mengekstrak informasi pada halaman web ke bentuk structured data *(record data)*.
2. Bagaimana menguji performansi dari sistem yang telah dibangun.

3. Apa kelebihan dan kekurangan metode ini.

Batasan masalah yg akan dibahas pada tugas akhir ini diantaranya:

1. Untuk kemudahan pada tahap pengujian, sistem ekstraksi halaman web akan diimplementasikan dalam bentuk sistem extraksi data produk dari 6 toko online.
2. Informasi yang akan diekstrak dari suatu halaman web hanya berupa text saja yaitu nama produk dan harga produk.

1.3 Tujuan

Tujuan dari tugas akhir ini adalah:

1. Mengimplementasikan metode *Vision Based Page Segmentation* untuk sistem pengekstrak informasi dari halaman web ke bentuk *structured data* (*record data).*
2. Menguji akurasi metode *Vision Based Page Segmentation* dengan menggunakan parameter *recall* dan *precission.*
3. Menganalisis sistem yang telah dibangun ini menitik beratkan pada *recall* atau *precision* dengan menggunakan parameter *harmonic mean* dan *the e measure*.

1.4 Metodologi penyelesaian masalah

1. Studi Literatur :
   Pengumpulan referensi berupa buku, jurnal, artikel, tutorial, dan referensi lain untuk mempelajari teori, konsep, dan informasi-informasi lain yang berkaitan dengan tugas akhir ini.
2. Analisis dan Desain :
   a) Menganalisis kebutuhan fungsionalitas serta *hardware* yang diperlukan untuk tugas akhir ini.
   b) Mendesain sistem yang akan dibangun dengan pemodelan UML (*Unified Modelling Language*).
   c) Mendesain skenario pengujian dan parameter-parameter yang akan digunakan pada saat pengujian.
3. Implementasi :
   Pengimplementasian ke bentuk aplikasi sehingga pengujian dapat dilakukan.
4. Pengujian

   Pengujian terhadap aplikasi yang telah dibuat sesuai dengan skenario dan parameter yang telah dirancang.
5. Analisis hasil :

   Analisis terhadap hasil pengujian yang telah dilakukan
6. Pengambilan kesimpulan dan penyusunan laporan tugas akhir :

   Pengambilan kesimpulan dari hasil analisis yang telah dilakukan pada tahap sebelumnya serta penyusunan laporan tugas akhir.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan analisis dan pengujian yang telah dilakukan terhadap sistem extraksi halaman web menggunakan metode *Vision Based Page Segmentation* (VIPS) ini, maka dapat diambil beberapa kesimpulan sebagai berikut:

1. Metode VIPS terbukti dapat digunakan untuk mengekstrak informasi yang diinginkan dari suatu halaman web.

2. Bagian paling kritikal dalam sistem ini adalah pada saat menentukan rule. Metode VIPS akan menghasilkan tingkat akurasi yang tinggi jika dapat didefinisikan rule-rule ekstraksi yang tepat.

3. Dari perhitungan *harmonic mean* dan *e-measure*, terlihat bahwa metode VIPS lebih menitikberatkan kepada *recall*, karena sistem ini lebih cenderung untuk mengambil semua data yang sesuai dengan pattern yang telah ditentukan.

4. Metode VIPS tidak menjamin data yang telah diekstrak telah memenuhi sifat *atomic value* untuk dimasukkan ke dalam database karena atomic atau tidaknya hasil ekstraksi akan tergantung pada desain dari halaman web yang diekstrak.

5. *Response time* bergantung pada besarnya ukuran *page* (banyaknya tag html) yang diekstrak, semakin besar ukuran page maka semakin besar pula *response time*-nya.

5.2 Saran

1. Penelitian lebih lanjut dengan menggabungkan metode web page extraction dengan metode web crawling.
2. Pengimplementasian metode web page extraction untuk membangun sistem yang siap pakai, misalnya: sistem pembanding harga produk antar toko online.