EKSTRAKSI TEKS PADA HALAMAN WEB BERITA MENGGUNAKAN WRAPPER INDUCTION
TEXT EXTRACTION FROM NEWS WEB PAGE USING WRAPPER INDUCTION

Mizana Khusnu Perdani[1], Arie Ardiyanti Suryani[2], Yanuar Firdaus A.w.[3]

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

---

Abstrak

Penggunaan Internet semakin pesat dan orang membutuhkan suatu cara untuk melihat content-content yang penting dari sebuah halaman Web. Hal inilah yang mendorong diciptakannya suatu teknologi untuk mengekstrak content atau informasi dari halaman Web tersebut sehingga dapat mempermudah dalam pembacaan dan analisis. Informasi pada halaman Web dapat berupa teks, gambar, alamat URL dan sebagainya. Karena bentuknya yang semi-structured, untuk mengambil informasi dari halaman Web cukup sulit.

Wrapper merupakan salah satu metode untuk mengekstrak halaman Web. Namun Wrapper mempunyai kelemahan, yaitu tidak adanya proses learning, sehingga sistem berjalan secara manual (hand coded), karena itulah dibuat suatu metode pengembangan dari Wrapper ini yang menyediakan proses learning yaitu Wrapper Induction. Proses learning pada Wrapper Induction ini adalah pada proses generate tag HTML sebagai penentu content-content yang akan diekstrak. Pada Tugas Akhir ini akan dilakukan ekstraksi informasi yang berupa teks berita menggunakan Wrapper Induction dan analisis perfomansi dari Wrapper Induction dalam mengekstrak halaman web berdasarkan Recall, Precision dan F-Measure.

Kata Kunci : Wrapper, Wrapper Induction, halaman Web

Abstract

The using of Internet is increase and people need a technique to get the important contents of a Web page. Because of that case, the technology to extract contents or information of a Web page had been invented, then the Web page can be both read and analyzed easily. Web page contains of many informations such as text, images, URL address and so on. Because of semi-structured, there‟s quite difficult to take information from Web Page.

Wrapper is a one of methods to extract a web page. But, Wrapper has a weakness; it doesn‟t have a learning process, then the system running manually (hand coded). Because of that case, Wrapper Induction which is provided a learning process had developed. Learning process on Wrapper Induction is a process to generate HTML tag to indentify which content will be extract. This Final Project is created to extract text information from news Web page using Wrapper Induction and analyze the performance of Wrapper Induction on extracting a Web page based on Recall, Precision and F-Measure.

Keywords : Wrapper, Wrapper Induction, Web page

---

1. Pendahuluan

1.1. Latar Belakang

Seiring dengan perkembangan jaman, manusia semakin dipermudah dengan adanya teknologi-teknologi yang berkembang. Dalam era globalisasi dan jaman serba cepat seperti sekarang, waktu menjadi sangat penting. Manusia membutuhkan sesuatu yang lebih praktis dan cepat. Internet contohnya, karena cepat dan mudah untuk diakses internet menjadi salah satu hal yang penting dan dibutuhkan manusia untuk mendapatkan berbagai informasi yang mereka butuhkan. *Website* tidak hanya mengandung content utama, tetapi juga mengandung *content* yang tidak berhubungan dengan content utama. Karena banyaknya *content* yang terkandung dalam website sedangkan manusia membutuhkan sesuatu yang lebih praktis, maka content-content tersebut harus diekstrak, sehingga informasi yang diperlukan dapat terlihat. Hasil ekstraksi ini juga dapat diintegrasikan dalam suatu data dan digunakan sesuai dengan kebutuhan, contohnya dalam perusahaan yang ingin membandingkan harga jual dengan perusahaan lain. Hal ini diharapkan akan mempercepat manusia dalam membaca dan menganalisis isi atau content dari suatu halaman web. Informasi pada halaman Web dapat berupa teks, image, alamat URL dan sebagainya. Tugas Akhir ini hanya akan membahas ekstraksi untuk informasi berupa teks pada halaman Web berita.

Banyak metode atau algoritma yang dikembangkan untuk mengekstrak halaman Web, salah satunya adalah dengan *Wrapper*. *Wrapper* adalah suatu teknik untuk mengekstrak informasi dari *semi-structured text* (contoh: HTML)[2]. *Wrapper* dikonstruksi secara manual untuk mengekstrak suatu informasi pada satu halaman Web serta tidak bisa digunakan untuk halaman Web lainnya, sehingga harus dilakukan konstruksi ulang pada setiap halaman Web yang akan diekstrak, karena itu *Wrapper* tidak praktis dan membutuhkan *cost* yang besar[7]. Beberapa sistem mengkombinasi teknik *Wrapper* tersebut agar lebih praktis dan efesien dengan mengkonstruksi *Wrapper* secara otomatis, salah satunya adalah dengan cara induksi dan pada Tugas Akhir ini akan digunakan metode *Wrapper Induction*. *Wrapper Induction* adalah suatu teknik yang secara otomatis akan mengkontruksi wrapper[6]. *Wrapper Induction* melakukan *learning* pada *training* data, dengan adanya *learning* tersebut, diharapkan waktu yang dibutuhkan untuk ekstraksi menjadi lebih sedikit karena hasil *learning* dapat digunakan untuk berbagai halaman Web. *Wrapper Induction* bersifat ekspresif dalam artian seberapa baik *Wrapper* dapat menangani sumber Internet yang aktual[6]. *Wrapper Induction* juga bersifat efesien, hal ini dapat diukur dengan jumlah halaman Web dan waktu yang dibutuhkan untuk belajar (*learning*). Proses learning pada *Wrapper Induction* tidak membuthkan banyak sample atau data training. Karena prinsip dari learning ini adalah efesiensi data training yaitu bagaimana menghasilkan suatu teks yang relevan dengan menggunakan data *training* yang tidak terlalu besar. Relevan yang dimaksud dalam Tugas Akhir ini adalah seluruh teks yang terekstrak mengandung informasi penting sesuai dengan berita pada halaman Web asli.

1.2. Perumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan sebelumnya, maka permasalahan yang diteliti antara lain:

1. Bagaimana cara untuk mengkstrak informasi teks pada halaman Web berita.
2. Bagaimana perfomansi dan akurasi algoritma *Wrapper Induction* dalam mengekstrak informasi teks halaman Web berita.

Adapun batasan masalah Tugas Akhir ini adalah sebagai berikut :

1. Output yang dihasilkan adalah hasil ekstraksi infotmasi teks dari halaman Web berupa sekumpulan paragraf.
2. *Content* halaman Web berformat HTML.
3. Website bersifat *offline.*
4. Informasi yang akan diekstrak berupa teks saja.
5. Halaman Web bersifat statis, dalam artian tidak selalu terupdate.
6. Web yang akan diekstrak adalah Web berita.
7. Akurasi sistem dihitung berdasarkan jumlah teks yang terekstrak pada satu halaman Web.

1.3. Tujuan

Tujuan yang ingin dicapai dalam penyusunan Tugas Akhir ini adalah sebagai berikut:

1. Mengimplementasikan ekstraksi teks pada halaman Web menggunakan algoritma *Wrapper Induction*.
2. Menganalisis akurasi ekstraksi teks pada halaman Web berdasarkan *precision* dan *recall*.
3. Menganalisis perfomansi algoritma *Wrapper Induction* berdasarkan pola delimiter yang terbentuk dari setiap jumlah halaman web yang dibutuhkan untuk proses *learning*.

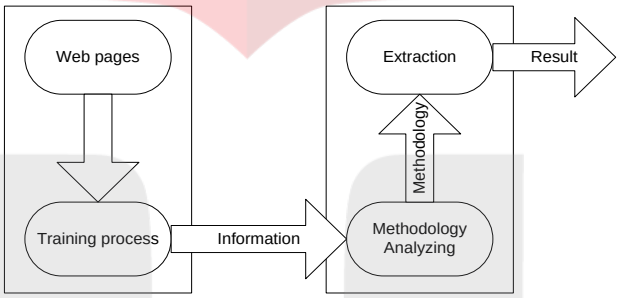
1.4. Metodologi Penyelesaian Masalah

Metodologi yang digunakan dalam memecahkan masalah di atas adalah dengan menggunakan langkah-langkah berikut:

1. Studi literatur
   Pencarian referensi dan sumber-sumber yang dapat digunakan sebagai acuan dalam pembuatan Tugas Akhir ini.
2. Analisis dan Design
   Adapun yang dianalisis adalah:
   a. Menganalisis data dan fungsionalitas yang akan diimplementasikan pada sistem
   b. Menganalisis metode *Wrapper Induction* yang akan digunakan dalam ekstraksi teks pada halaman web
   c. Menganalisis parameter yang akan digunakan untuk mengukur akurasi.

   Yang akan di design dalam sistem ini adalah:
   a. Design modul yang akan digunakan
   b. Design database
   c. Design fungsi dan prosedur sistem
   d. Design interface

3. Implementasi sistem
   Membangun perangkat lunak yang dapat mengekstraksi informasi teks dari halaman web berita menggunakan Wrapper Induction. Adapun garis besarnya adalah:
   a. Menentukan information resource yaitu halaman Web yang tersedia pada resource di internet, pada Tugas Akhir ini Web yang akan digunakan adalah Web berita.
   b. Menampilkan web pages.
   c. *Training* data (proses *learning*).
   d. Menganalisis metodologi yang digunakan yaitu wrapper induction.
   e. Melakukan ekstraksi halaman web.
      Adapun gambar prosesnya adalah sebagai berikut:



Gambar 1-1 : Proses Ekstraksi Halaman Web

4. Testing
   Menguji sistem termasuk menguji keakuratan Wrapper Induction dalam mengkstrak teks yang ada dalam halaman Web dan menguji ketepatan informasi dari hasil ekstraksi.
5. Analisis hasil
   Menganalisis hasil dari implementasi ekstraksi teks pada halaman Web berdasarkan precision dan recall serta menganalisis perfomansi algoritma *Wrapper Induction* berdasarkan pola delimiter yang terbentuk dari setiap jumlah halaman web yang dibutuhkan untuk proses *learning*.
6. Pengambilan kesimpulan
   Mengambil kesimpulan dari analisis pada point sebelumnya.
7. Pembuatan Laporan
   Membuat laporan yang mencakup seluruh penelitian yang telah dilakukan.

5. Kesimpulan dan Saran

5.1. Kesimpulan

Berdasarkan analisis dan pengujian terhadap algoritma *Wrapper Induction* yang telah dilakukan maka dapat diambil kesimpulan sebagai berikut:

1. Pembentukan pola delimiter pada proses *learning* tidak terlepas dari faktor content Web *training* selain jumlah halaman yang di-*training.*
2. Munculnya tag baru pada *learning* di mana tag tersebut jarang muncul di setiap halaman Web sangat mempengaruhi hasil ekstraksi, karena pola delimiter yang terbentuk tidak mampu untuk mengekstrak semua halaman Web dengan baik.
3. Ada kasus dimana pola delimiter yang dihasilkan dari *learning* suatu Web tidak dapat mengekstrak halaman Web lain dengan baik karena faktor content dan karakteristik Web tersebut dan hanya beberapa Web yang mampu diekstrak dengan menghasilkan akurasi yang cukup tinggi.
4. Akurasi dari pengujian menggunakan halaman yang baru dengan data uji 2 menghasilkan akurasi yang tidak jauh berbeda dengan pengujian menggunakan data uji 1, karena content halaman Web yang sama.
5. Pada pengujian menggunakan data uji 3, content dari data uji dan pola delimiter yang dihasilkan oleh data latih sangat mempengaruhi akurasi, jika delimiter cocok dengan content data uji maka akurasi akan semakin baik

5.2. Saran

1. Iterasi pada proses *learning* dapat ditambahkan sehingga pola delimiter yang dihasilkan dapat lebih adaptive.
2. Dibutuhkan pengembangan *Wrapper Induction* agar bisa bekerja pada sintaks selain HTML yaitu CSS dan Java Script
3. Dibutuhkan pengembangan *Wrapper Induction* agar tidak hanya dapat mengekstrak teks saja, namun juga gambar atau image.
4. Pengambilan halaman web lebih baik jika dilakukan secara online.
5. Ekstraksi informasi berupa teks ini dapat dikembangkan lagi untuk ekstraksi inti dari teks tersebut.

Daftar Pustaka

[1] Cowie, Jim., and Wilks, Yorick., *Information Extraction.*

[2] Chulkov, Georgi., 2000, *Wrapper Induction LR Wrapper*, Jacobs University Bremen.

[3] Even-Zohar, Yair., 2002, *Introduction to Text Mining*, National Center for Supercomputing Application University of Illinois.

[4] Gaizaukas, Robert., *An Information Extraction Perspective on Text Mining: Tasks, Technologies and Prototype Applications*, Natural Language Processing Group, Department of Computer Science, University of Sheffield.

[5] Grishman, Ralph., *Information Extraction: Techniques and Challenges*, Computer Science Department, New York University, New York, NY 10003, U.S.A.

[6] Kushmerick, Nicholas., 1997, *Wrapper Induction for Information Extraction*, University of Washington.

[7] Lam, Man I., Gong, Zhiguo., and Muyeba, Maybin., *A Method for Web Information Extraction*, Faculty of Science and Technology University of Macau.

[8] *Learning Wrappers Efficiently for Web Information Extraction Using Unlabeled Examples*

[9] McCluskey, Lee, *Information Extraction from the WWW using Machine Learning Techniques*, Dept of Informatic.

[10] Mooney, Raymond J., and Bunescu, Razvan., *Mining Knowledge from Text Using Information Extraction*, Department of Computer Sciences, University of Texas at Austin.

[11] Qu, Hongfei., 2001, *Wrapper Induction: Construct wrappers automatically to extract information from web sources Computing Science Department*, Simon Fraser University.

[12] Bing Liu, 2005, *Web Content Mining*, University of Illinois at Chicago.

[13] Dr. Spiros Sirmakesis, 2003, *Web Mining Past, Present and Future*, Computer Technology Institute.