

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 efisien 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 mengkonstruksi 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 efisien, hal ini dapat diukur dengan jumlah halaman Web dan waktu yang dibutuhkan untuk belajar (*learning*). Proses *learning* pada *Wrapper Induction* tidak membutuhkan banyak sample atau data training. Karena prinsip dari *learning* ini adalah efisiensi 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 mengekstrak informasi teks pada halaman Web berita.
2. Bagaimana performansi 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 informasi 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 performansi 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:

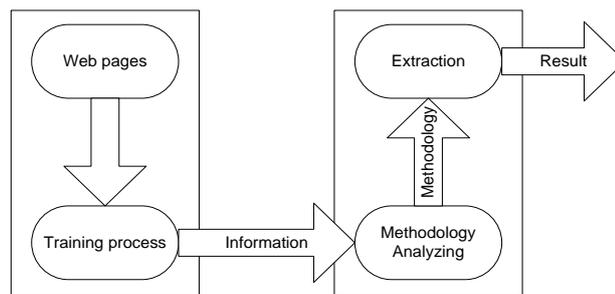
1. Studi literatur
Pencarian referensi dan sumber-sumber yang dapat digunakan sebagai acuan dalam pembuatan Tugas Akhir ini.
2. Analisis dan Design
Adapun yang dianalisis adalah:
 - a. Menganalisis data dan fungsionalitas yang akan diimplementasikan pada sistem
 - b. Menganalisis metode *Wrapper Induction*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