1. Pendahuluan

1.1. Latar belakang

Seiring dengan pesatnya pertumbuhan sumber informasi yang tersedia di *World Wide Web* pada jaringan internet, bermunculan juga teknik-teknik *data mining* yang ditujukan untuk mempermudah pemakai dalam memperoleh data yang ada sesuai dengan keinginannya. Akan tetapi terdapat permasalahan untuk mengidentifikasi potongan informasi relevan pada halaman web karena sering dikacaukan dengan isi tidak relevan seperti iklan, *navigation-panel*, *copyright notice* dan lain-lain disekeliling inti isi dari halaman web. Diperlukan suatu alat untuk mengekstrak informasi dari halaman web untuk menyediakan nilai tambah layanan, yaitu mengekstrak data(*instance*) dalam daerah khusus(*template*) pada halaman web yang sesuai dengan keinginan user.

Ada beberapa teknik yang sudah ada untuk mengekstrak data dari halaman web termasuk teknik manual dan teknik otomatis. Pendekatan manual untuk mengeluarkan data dari sumber-sumber web adalah dengan cara menulis program-program khusus, yang disebut *wrappers*, bahwa data mengidentifikasi hal penting dan memetakan mereka untuk beberapa format yang sesuai. Mengembangkan *wrappers* secara manual memiliki banyak kelemahan, terutama karena kesulitan dalam penulisan dan pemeliharaan *wrappers*.

Teknik otomatis atau sering disebut dengan *Automatic Data Extraction* merupakan cara lain untuk mengekstrak data. Omini [1], Information Extraction Based on Pattern Discovery(IEPAD) [4] dan Mining Data Record(MDR) merupakan contoh algoritma dari metode *Automatic Data Extraction*. Ketiga algoritma tersebut membentuk *pattern* untuk ekstraksi data. Akan tetapi *pattern* yang dihasilkan tidak spesifik karena semuanya ditentukan oleh sistem. Oleh karena itu ekstraksi data yang dihasilkan belum tentu sesuai dengan keinginan user.

Instace-Based Dara Extraction (IDE) merupakan salah satu dari sekian banyak algoritma pada metode *Automatic Data Extraction*. Algoritma ini masuk dalam kategori *Automatic Data Extraction* karena mencari *pattern* yang berulang kemudian menggunakan *pattern* tersebut untuk mengkestrak data. Diharapkan dengan adanya keterlibatan user dalam pemilihan data yang akan diekstrak memberikan pengaruh yang signifikan terhadap hasil ekstraksi.

Dalam penyusunan Tugas Akhir ini akan dilakukan implementasi metode ADE menggunakan algoritma IDE. Dalam proses ekstraksi data, metode ini melibatkan user dalam memberikan label pada halaman web sehingga dapat memberikan hasil yang sesuai dengan keinginan user. Algoritma IDE hanya membutuhkan satu halaman web untuk membuat *pattern* yang akan digunakan untuk mengkestrak data.

1.2. Perumusan Masalah

Berdasarkan latar belakang tersebut, maka permasalahan yang akan dibahas adalah sebagai berikut :

1. Bagaimana mengimplementasikan metode *Automatic Data Extraction* dengan algoritma IDE untuk ekstraksi data pada halaman web.

2. Bagaimana menentukan *pattern* yang dapat secara tepat mengekstraksi data sehingga sesuai dengan keinginan user.
3. Apa saja faktor yang mempengaruhi performansi algoritma IDE.
4. Bagaimana melakukan pengukuran dan pengujian untuk mengetahui performansi dari sistem IDE yang dibangun.

Batasan masalah untuk penelitian ini adalah :

1. Dokumen web yang digunakkan pada penelitian ini terbatas pada dokumen HTML.
2. Kategori dokumen web terbatas pada website jenis daftar produk dan jenis berita.
3. Implementasi yang dibuat dalam tugas akhir ini menekankan hanya pada proses mendapatkan inti informasi dari sebuah halaman web yang berupa teks dan image, jadi tidak termasuk link ke halaman lain.
4. Item yang diekstrak terbatas pada pilihan yang disediakan oleh aplikasi.
5. Yang diamati hanyalah performansi yang berupa *recall ratio* dan *precission ratio*.

1.3. Tujuan

Tujuan dari dilakukannya penelitian ini adalah:

1. Mengetahui ketepatan *pattern* yang di bentuk dengan menggunakan metode *Automatic Data Extraction* dengan algoritma IDE.
2. Menghasilkan ekstraksi data sesuai keinginan user.
3. Menganalisa faktor-faktor yang mempengaruhi performansi berupa *range node* dan jenis website inputan.
4. Menganalisa sistem yang telah dibangun ini dilihat dari *recall ratio* dan *precision ratio*.

1.4. Metodologi penyelesaian masalah

Penelitian ini akan mengimplementasikan ekstraksi *Web Content* dengan *Automatic Data Extraction* menggunakan metode *Instance-based Data Extraction*.

1. Studi Literatur :

Pencarian referensi dan sumber-sumber lain yang dapat digunakan sebagai acuan dalam penelitian dan pembelajaran ekstraksi pada *Web Content*. Meliputi tentang proses pemberian label pada web, ektraksi struktur data, dan pencocokan tag pada source HTML.

2. Pengumpulan data :

Mencari studi kasus website yang akan digunakan untuk sumber data.

3. Analisis dan Desain :

Tahapan ini adalah tahapan yang meliputi analisis terhadap *Instance-based Data Extraction* kemudian akan dilakukan perancangan aplikasi untuk mengekstraksi data dari halaman web.

4. Implementasi :

Tahap ini adalah tahap pembangunan aplikasi. Pada tahap ini dibangun perangkat lunak yang akan digunakan sebagai antar muka untuk melakukan proses ekstraksi pada konten halaman web.

5. Testing :

Pada tahap ini akan dilakukan pengujian terhadap perangkat lunak yang telah dibangun, apakah sudah bekerja dengan benar atau belum. Pengujian dilakukan dengan mensimulasikan proses ekstraksi data pada suatu website.

6. Analisis hasil :

Analisis faktor-faktor yang mempengaruhi proses ekstraksi oleh *Instance-based Data Extraction* dari website.

7. Pengambilan kesimpulan dan penyusunan laporan tugas akhir :

Pengambilan kesimpulan dari hasil analisis yang telah dilakukan pada tahap sebelumnya untuk kemudian disusun laporan terhadap analisis yang telah dilakukan.