1. Pendahuluan

1.1 Latar belakang

Seiring dengan pesatnya pertumbuhan sumber informasi yang tersedia di *World Wide Web* pada jaringan internet, kemudian bermunculan teknik-teknik *data mining* yang ditujukan untuk mempermudah pemakai dalam memperoleh data yang ada sesuai dengan keinginannya. Untuk memperoleh fitur-fitur dari produk di berbagai situs web, maka itu sangat susah dilakukan untuk mengidentifikasi potongan informasi relevan karena halaman web sering dikacaukan dengan isi tidak relevan seperti iklan, *navigation-panel*, *copyright notice* dan lain-lain disekeliling inti isi dari halaman web. Sehingga diperlukan suatu alat untuk mengekstrak informasi dari halaman web untuk menyediakan nilai tambah layanan, yaitu mengekstrak deskripsi atribut dari setiap produk (*data obyek*) dalam daerah khusus (*data region*) pada halaman.

Data record adalah informasi dalam jumlah besar pada web diisi di struktur obyek yang teratur. *Data record* begitu penting karena sering menampilkan inti informasi dari halaman *host*-nya. Selain itu, sebuah daftar berupa obyek dalam halaman web sering mendeskripsikan sebuah daftar barang yang serupa, contohnya daftar produk atau layanan. Dengan begitu dapat disebut database dari record-record yang ditampilkan pada halaman web dengan pola teratur. *Data record* pada halaman web yang di-*mining* dapat bermanfaat karena dapat digunakan untuk mengekstrak dan mengintegrasikan informasi dari berbagai macam sumber untuk menyediakan nilai tambah layanan, contohnya, pengumpulan informasi web sesuai dengan keinginan (*customizable web information gathering*), perbandingan harga (*comparative-shopping*), *metasearch* dan lain sebagainya.

Ada beberapa teknik yang sudah ada untuk me-*mining data record* dari halaman web termasuk teknik manual, *supervised learning*, dan teknik otomatis. Pada teknik manual, seorang programmer terlebih dahulu harus mengamati sebuah halaman web dan *source code*-nya, setelah menemukan beberapa pola dan kemudian baru bisa menulis program untuk mengindentifikasi setiap *data record*. Kelemahan metode manual ini adalah tidak mampu dalam menangani halaman web dalam jumlah yang besar. *Supervised learning* memerlukan secara manual menyiapkan data training positif maupun negatif dan juga masih memerlukan bagian kecil dari usaha manusia yaitu pelabelan bagian secara spesifik pada halaman web untuk menandai letak *data record*. Sedangkan teknik otomatis yang sudah ada, seperti OMINI [1] dan IEPAD [3] masih belum memuaskan karena kemampuannya yang kurang baik. OMINI merupakan metode yang dibangun dengan beberapa heuristik seperti *sibling tag heuristic* yaitu menghitung jumlah pasangan dari tag sibling pada tag tree, dan *partial path heuristic* yaitu daftar lintasan dari sebuah node untuk mencapai ke semua node lainnya serta jumlah kejadian pada setiap lintasan. IEPAD merupakan metode otomatis lainnya yang bertujuan untuk menemukan pola-pola dari tag string pada HTML dan kemudian menggunakan pola-pola tersebut dalam mengekstrak obyek. Metode IEPAD ini menggunakan *PAT tree* (sebuah *Patricia tree*) untuk menemukan pola. Masalah pada *PAT tree* adalah hanya mampu untuk menemukan obyek jika mempunyai pola sesuai dengan pola pada *PAT tree* saja, sehingga sering menghasilkan banyak

pola tetapi kebanyakan *data record* yang dihasilkannya adalah palsu bahkan *data record* sesungguhnya tidak ditemukan.

Pendekatan yang dilakukan pada tugas akhir ini adalah mengimplementasikan algoritma yang disebut MDR (*Mining Data Records in Web Pages*) secara otomatis dalam me-*mining data record* pada halaman web. Metode ini lebih efektif dalam melakukan ekstraksi informasi karena hanya berdasarkan pada dua pengamatan, yaitu mengamati *data record* pada halaman web dan algoritma pencocokan string.

1.2 Perumusan Masalah

Dalam tugas akhir ini, terdapat beberapa permasalahan yang timbul selama proses untuk *mining data record* pada halaman web diantaranya :

1. Bagaimana mengimplementasikan algoritma MDR dalam memperoleh *data record* secara otomatis pada suatu halaman web.
2. Bagaimana menentukan nilai batasan *edit distance* dengan menggunakan hanya beberapa halaman web dari dataset, dimana nilai yang terpilih itu akan dipakai pada semua pengujian nantinya.
3. Bagaimana melakukan pengukuran dan pengujian untuk mengetahui performansi dari sistem MDR yang dibangun.
4. Bagaimana melakukan perhitungan untuk mengetahui sistem yang telah dibangun ini menitik beratkan pada recall atau precision.

1.3 Tujuan

Tujuan pembahasan dari tugas akhir ini adalah :

1. Dengan nilai batasan *edit distance* yang terpilih, akan diuji prosentase akurasi sistem MDR dengan menggunakan parameter recall dan precision.
2. Menganalisia sistem yang telah dibangun ini menitik beratkan pada *recall* atau *precision* dengan menggunakan parameter the harmonic mean, dan the e measure.

1.4 Batasan Masalah

Batasan masalah untuk tugas akhir ini adalah sebagai berikut :

1. Tidak melakukan evaluasi dengan menggunakan banyak halaman dari satu situs tertentu karena kebanyakan halaman web dalam satu situs yang sama mempunyai pola yang serupa. Hal ini lebih berguna untuk menguji sebuah halaman dari berbagai situs dalam jumlah yang banyak.
2. Implementasi yang dibuat dalam tugas akhir ini menekankan hanya pada proses mendapatkan inti informasi dari sebuah halaman web yang berupa teks dan image (dalam hal ini akan memanfaatkan tag string HTML), jadi tidak termasuk link ke halaman lain.
3. Implementasi yang dibuat dalam tugas akhir ini hanya mengkosentrasikan pembangunan *HTML tag tree* dimulai dari tag <body> untuk mempersingkat waktu parsing.

1.5 Metodologi penyelesaian masalah

Metodologi yang akan digunakan dalam penyelesaian tugas akhir ini adalah :

1. Studi pustaka

Tahapan untuk menambah wawasan dari buku-buku, artikel dan sumber-sumber lain yang layak, seperti informasi-informasi yang tersedia di internet untuk menunjang pembahasan tugas akhir ini.

2. Analisis dan perancangan.
   Tahapan untuk menentukan kebutuhan sistem, seperti identifikasi input, identifikasi output, identifikasi spesifikasi hardware maupun perancangan software yang akan dibangun dalam bentuk diagram yang akan memudahkan pemahaman terhadap perangkat lunak tersebut.
3. Implementasi perangkat lunak
   Tahapan untuk menentukan nilai batasan *edit distance* serta implementasi dari *mining data record* pada halaman web dengan menggunakan algoritma MDR.
4. Pengujian perangkat lunak
   Menguji perangkat lunak yang dihasilkan sekaligus menganalisis hasil uji coba perangkat lunak yang telah dibangun.
5. Pengambilan kesimpulan dan penyusunan makalah
   Mengambil kesimpulan dari hasil pengujian dan pengukuran yang dilakukan serta menyusun makalah.

1.6 Sistematika Penulisan

Tugas akhir ini disusun berdasarkan sistematika sebagai berikut :

Bab I : Pendahuluan
Bab ini akan membahas kerangka penelitian atau percobaan dalam tugas akhir, meliputi latar belakang masalah, perumusan masalah, tujuan, batasan masalah, metode penyelesaian masalah, dan sistematika penulisan.

Bab II : Dasar Teori
Bab ini memuat berbagai dasar teori yang mendukung dan mendasari penulisan tugas akhir ini, yaitu mengenai konsep dari *web mining, data Mining* dan *clustering,* konsep *mining data record,* recall dan precision, the harmonic mean, dan the e measure.

Bab III : Analisis dan Perancangan Sistem
Berisi analisis sistem MDR yang akan dibuat mencakup analisis kebutuhan sistem, perancangan proses dan aliran data sehingga proses dapat dipahami secara jelas

Bab IV : Pengujian
Berisi tentang hasil pengujian sistem MDR yang telah dibuat.

Bab V : Kesimpulan dan Saran
Berisi tentang kesimpulan dari keseluruhan aplikasi yang dibuat serta saran untuk pengembangan selanjutnya.