1 Pendahuluan

1.1 Latar Belakang Masalah

Sejak lahirnya Internet, informasi dalam web berkembang secara pesat. Masyarakat yang sebelumnya menggunakan media konvensional sebagai sarana publikasi informasi seperti surat kabar, majalah, tabloid, pamflet dan lain sebagainya mulai beralih memanfaatkan internet karena dirasa lebih efisien dalam mempublikasikan informasi yang mereka hasilkan ataupun untuk mencari informasi yang mereka kehendaki. Sebagai akibatnya informasi yang beredar di Internet terus meningkat secara eksponensial.

Informasi yang ditampilkan dalam internet biasanya berupa halaman web yang berformat HTML. Dalam suatu halaman web terdapat berbagai macam informasi yang dapat ditampilkan, walaupun sebenarnya hanya sebagian kecil dari halaman tersebut inti informasi yang ingin disampaikan. Ada berbagai informasi tambahan dari suatu halaman web yang tidak ada hubungannya dengan konten utama dari web tersebut yang disebut dengan *noise*. Informasi tambahan tersebut diantaranya adalah panel navigasi, *event*, *related links*, *copyright*, sinopsis suatu berita, berbagai macam iklan dan lain-lain yang secara keseluruhan bertujuan untuk mempermudah pengguna dalam mengakses informasi dalam halaman web tersebut.

Saat ini para pengguna Internet yang akan mengakses informasi lebih dimudahkan dengan adanya layanan *search engine* yang beragam jenisnya, salah satunya adalah *news search engine* yang memiliki fokus utama pada halaman web berita. Dalam sistem search engine ada yang dinamakan dengan proses *indexing* yang berfungsi untuk untuk mengoptimalkan kecepatan dan kinerja dalam menemukan dokumen yang relevan untuk permintaan pencarian. Tanpa *index*, *search engine* akan memeriksa setiap dokumen dalam *document collection*, yang akan memerlukan banyak waktu dan daya komputasi. Secara logika hasil dari suatu *query* dari *search engine* akan lebih baik jika proses *indexing* yang dilakukan dengan meng-*index content* saja dari suatu halaman website. Oleh sebab itu, diperlukan suatu proses yang dapat memisahkan *content* utama halaman web dengan *noise*, proses ini disebut dengan *web extraction* yang menjadi fokus utama dalam tugas akhir ini.

Selain itu, halaman web yang ditampilkan di Internet tidak semuanya memenuhi standar W3C (World Wide Web Consortium). Masih banyak halaman web memiliki struktur tag yang belum valid, oleh sebab itu nantinya sistem *web extraction* yang dibangun akan coba diujikan pada dua jenis dataset yaitu mentah dan valid untuk mengetahui pengaruhnya pada performansi sistem.

Adapun dalam tugas akhir ini metode *web extraction* yang digunakan untuk mengambil informasi utama dari halaman web berita adalah dengan metode *hybrid* yang merupakan kombinasi atau penggabungan dari konsep *tree* dan konsep *tag sequence*.

1.2 Perumusan Masalah

Berdasarkan latar belakang diatas, maka permasalahan yang akan diangkat dalam penelitian ini adalah sebagai berikut :

1. Bagaimana mendeteksi dan mengeliminasi *noise* yang ada dalam suatu halaman web sehingga didapatkan informasi utama dari website tersebut.
2. Bagaimana mengevaluasi proses *deteksi* dan *eliminasi* yang dilakukan dengan menghitung precision, recall dan f-measure.

1.3 Tujuan

Berdasarkan rumusan masalah diatas, tujuan yang ingin dicapai dari tugas akhir ini adalah :

1. Mengimplementasikan ekstraksi informasi utama halaman web berita dengan menggunakan metode *hybrid*.
2. Mengetahui pengaruh penggunaan jumlah dataset yang digunakan dengan performansi yang ditunjukkan.
3. Mengetahui perbedaan performansi dataset mentah dengan dataset yang telah divalidasi dengan validator W3C berdasarkan parameter precision, recall dan f-measure.

1.4 Batasan Masalah

Batasan masalah untuk tugas akhir ini adalah sebagai berikut :

1. Menggunakan metode hybrid untuk mendapatkan informasi utama dalam halaman web berita.
2. Halaman website yang dijadikan dataset adalah halaman yang berisi berita lengkap, bukan synopsis atau berupa headline saja (halaman index tidak dijadikan dataset).
3. Dataset yang diambil dalam bentuk HTML pages secara offline hasil dari proses *crawling* halaman website berita yang telah ditentukan dengan menggunakan tools teleport.
4. Tiap dataset dari masing-masing situs dikelompokkan dalam direktori yang memiliki nama sesuai dengan nama situsnya untuk memudahkan identifikasi, dan masing-masing direktori memiliki 3 subdirektori yaitu mentah untuk menyimpan data mentah, valid untuk menyimpan data yang telah divalidasi dan real content untuk menyimpan file teks yang berisi content sebenarnya dari suatu halaman website.

1.5 Metodologi Penyelesaian Masalah

Penelitian ini akan mengimplementasikan proses ekstraksi data dari website berita. Parameter yang akan dilihat adalah precision, recall dan F-measure dari metode yang digunakan.

1. Studi Literatur :
   Mencari referensi yang layak dan berhubungan dengan topik yang diangkat, memahami dan mempelajari tentang *text mining*, *web mining*, *web page cleaning / web extraction*, teknik eliminasi *web pages Noise* dari berbagai jurnal, buku, Internet dan referensi lainnya yang mendukung
2. Pengumpulan dan pengolahan data :

Melakukan pengumpulan artikel berita yang berbentuk html yang akan digunakan sebagai dataset.

3. Analisis dan Desain :
   Identifikasi artikel-artikel berita utama dalam halaman web tersebut (*web data extraction*) dengan menghilangkan semua informasi yang tidak terkait (*noise*) dan kemudian dilakukan proses *eliminasi* untuk mendapatkan pola informasi berita dari struktur HTML web pages.
4. Implementasi dan Pengujian :
   - Implementasi perangkat lunak
     Mengimplementasikan perancangan menjadi perangkat lunak. Proses implementasi perangkat lunak dilakukan berdasarkan dari proses analisis dan perancangan yang telah dibangun.
   - Pengujian
     Memeriksa error handling yang ada pada perangkat lunak yang dibangun misalnya kesalahan perhitungan, kesalahan dalam penginputan data, human error dan lain sebagainya.
5. Analisis hasil :
   Melakukan analisis dari hasil yang telah diperoleh dengan menghitung nilai precision, recall dan F-Measure.
6. 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.

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 page noise, web Mining,* konsep *web extraction,* recall dan precision, serta F-Measure.

Bab III : Analisis dan Perancangan Sistem

Berisi analisis sistem Web Extraction yang akan dibuat mencakup analisis kebutuhan sistem, perancangan diagram use case dan diagram kelas.

Bab IV : Implementasi dan Pengujian

Berisi tentang hasil pengujian sistem Web Extraction yang telah dibuat.

Bab V : Kesimpulan dan Saran

Berisi tentang kesimpulan dari keseluruhan aplikasi yang dibuat serta saran untuk pengembangan selanjutnya.