BAB I
PENDAHULUAN

1.1 LATAR BELAKANG

Berkembang pesatnya dunia internet dan kebebasan dari sesorang untuk membuat suatu halaman web maka mengakibatkan halaman web berkembang jumlahnya dengan sangat pesat dan hal tersebut menjadi suatu permasalahan untuk seseorang melakukan pencarian data yang memang dibutuhkan dari suatu halaman web. Suatu web juga dapat dikatakan kaya apabila konten-konten yang ada pada halaman web tersebut dapat dilihat ataupun tersedia sehingga terdapat akses terhadap konten tersebut oleh pengguna. Konten pada halaman web bisa terdiri dari artikel maupun sebuah dokumen yang diunggah pada halaman web tersebut. Pemindaian suatu halaman web dapat difokuskan dengan mencari dokumen apa saja yang tersimpan dalam web tersebut yang dapat diakses oleh publik.

Crawler adalah suatu program yang di pergunakan untuk mengambil file suatu situs. Dalam prakteknya banyak pemilik situs atau webmaster mendaftarkan situsnya ke search engine seperti Google, Yahoo!, HotBot, Alta Vista, Web Crawler, Infoseek, dan Lycos. Search Engine ini secara berkala akan mengakses situs terdaftar tersebut kemudian menarik data (data mining) teks html dalam situs tersebut untuk dilakukan pengindeksian atas isi situs, pengindeksian ini dilakukan untuk memberi gambaran mengenai isi situs tersebut kepada orang yang melakukan pencarian melalui search engine.

Berdasarkan uraian latar belakang diatas, maka penulis tertarik untuk mengambil tugas akhir dengan judul **"Membangun Web Crawler Untuk Pengindeksian Data Teks Pada Web Koran Online Di Aceh"**

1.2 RUMUSAN MASALAH

Berdasarkan uraian di atas, maka permasalahan yang timbul dalam pengerjaan tugas akhir ini adalah:

1. Bagaimana mengembangkan aplikasi Web Crawler untuk pengindeksian data teks.
2. Bagaimana membangun sebuah aplikasi yang akan mengambil semua file web dari sebuah situs.
3. Bagaimana menyimpan konten dari web yang di ambil kedalam file text .txt

1.3 BATASAN MASALAH

Dalam melakukan penelitian ini sangat diperlukan adanya batasan-batasan yang ditetapkan agar lebih terfokus dan tidak melebar. Adapun batasan-batasan yang diperlukan yaitu:

1. Sistem hanya akan mengambil file yang ekstensi html, htm, php, asp
2. File teks yang disimpan hanya berisikan kata dasar dengan membuang semua tag html dan tag php serta tag asp dari file yang di crawler.
3. Inputan pada sistem yaitu alamat web dari koran online yang ada di Aceh.

1.4 TUJUAN PENELITIAN

Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Membangun sebuah aplikasi web crawler untuk mengambil isi dari laman web berita online di Aceh.
2. Menyimpan semua kata dasar dari isi laman web menjadi sebuah file text.

1.5 RELEVASNI

Setelah Web Cralwer ini selesai, diharapkan web cralwer ini bisa menampilkan file teks dari sebuah situs yang bisa memudahkan seseorang dalam

mencari dokumen yang dibutuhkan berdasarkan keyword atau kata kunci yang tepat.