BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Teknologi Informasi saat ini mengalami perkembangan yang signifikan. Beragam aspek kehidupan sangat terbantu dengan perkembangan teknologi informasi ini. Hal ini dirasakan oleh berbagai pihak yang mulai merasa bahwa teknologi informasi adalah hal yang sulit dipisahkan dari pekerjaan sehari-hari.

Semakin menjamurnya teknologi *internet* ternyata justru memberikan berbagai dampak positif maupun negatif pada penggunanya. Hal ini disebabkan informasi mudah didapat dan juga mudah pula disebar. *Internet* yang menjadi salah satu idola baru dalam pencarian dan penyebaran informasi inipun memang sangat menjanjikan, karena teknologi ini secara statistik perkembangannya sangat mengagumkan. Pada akhir tahun 2011, Uni Telekomunikasi Internasional mengumumkan bahwa pengguna *internet* di dunia secara global menembus angka 2.095.006.005. (internetworldstats.com, 2011)

Jumlah *website* yang berisikan data dan informasi jumlahnya selalu meningkat setiap harinya dikarenakan faktor jumlah pengguna yang aktif sangatlah banyak. Terbukti menurut data stastistik yang dikeluarkan pihak netcraft bahwa pada akhir tahun 2010 jumlah *website* di *internet* mencapai 255.287.546 *website*. Dengan banyaknya jumlah *website* ini para pengguna *internet* pun membutuhkan bantuan dari *search engine* untuk mencari informasi *website* yang diinginkan. (netcraft.com, 2010)

Contoh kasus yang sering muncul misalnya saat pengguna internet ingin mengakses informasi tentang *office*. *Office* yang dimaksud disini adalah tentang perkantoran. Masalah yang muncul adalah halaman website yang mengandung kata *office* berjumlah 1.960.000.000, dengan halaman pertama hasil pencarian bukanlah tentang perkantoran akan tetapi justru perangkat lunak *Microsoft Office*. Kondisi yang terjadi sangat wajar karena *Microsoft Office* adalah produk yang paling banyak digunakan dan dicari. Oleh karena itu, berdasarkan banyaknya halaman website yang mengandung *keyword office*, pengguna pasti kebingungan mencari informasi yang sesuai dengan kebutuhannya. (google.com, 2011)

Search engine Google pada kasus ini melakukan pengurutan hasil pencarian dari halaman sebuah *website* untuk mengukur tingkat relevansi sebuah halaman dengan suatu *keyword* tertentu dengan cara memperhatikan beberapa faktor yang diperhatikan dalam formula algoritma *Page Rank* yang digunakan oleh Google. Berikut adalah formula perhitungan *Page Rank* yang diterbitkan oleh Sergey Brin dan Lawrence Page dalam jurnalnya yang berjudul *The Anatomy of a Large-Scale Hypertextual Web Search Engine* :

```
PR(A) = (1-d) / N + d ( ( PR(T1) / C(T1) )+ … +( PR(Tn) / C(Tn)))
```

Faktor-faktor yang diperhatikan oleh Google terhadap pengurutan hasil pencarian dari sebuah keyword terdiri dari :

a. PR(A) yaitu *Pagerank* halaman A (halaman yang saat ini sedang dikunjungi).

b. PR(T1) yaitu *Pagerank* halaman *T1* (halaman lain) yang mengacu ke halaman A(halaman yang saat ini sedang dikunjungi).

c. C(T1) yaitu jumlah link keluar (*outbound link*) pada halaman T1.

d. d yaitu *Damping factor* yang bisa diberi nilai antara 0 dan 1. Damping factor dapat diartikan juga sebagai peluang link tersebut akan dikunjungi atau diikuti oleh pengunjung web tersebut.

e. N yaitu jumlah keseluruhan halaman *web* (yang terindex google).

Untuk membantu pengguna dalam mencari informasi yang relevan, *search engine* tentu terlebih dahulu harus memiliki database tentang halaman *website* yang ada di internet. Ada dua cara search engine untuk mengumpulkan data halaman *website* di internet. Pertama dengan cara konvensional yakni *search engine* melakukan pengunjungan halaman-halaman *website* yang ada di internet menggunakan *web crawler*, dan cara yang kedua dengan menyediakan *tools* optimisasi untuk para webmaster memasukan informasi tentang websitenya.

Cara yang pertama mempunyai tahapan-tahapan sebagai berikut :

1. *Web Crawler* sebagai mesin yang mempunyai tugas akan menelusuri setiap website yang ada di internet.
2. Saat mengunjungi halaman website *Web Crawler* akan mengumpulkan semua link yang ada dengan cara melakukan filtering konten menggunakan url recognizer.
3. Setelah itu *Web Crawler* akan menyimpan atribut yang ada pada *title* dan *meta tag* yang berfungsi untuk menentukan relevansi *keyword* dengan konten.

4. Tahap terakhir *Web Crawler* akan mengunjungi halaman lainnya dari link yang telah didapatkan sebelumnya untuk kemudian datanya disimpan kedalam database mesin pencari.

Cara yang kedua adalah *search engine* menyediakan fasilitas yang dapat digunakan oleh para webmaster untuk memasukan informasi websitenya. Cara kedua ini dapat mempercepat dikenalinya sebuah *website* atau halaman *website* oleh *search engine* karena beberapa informasinya akan dibantu dipenuhi oleh para pemilik *website* tersebut, sehingga beban *search engine* hanya tinggal melakukan *crawling* informasi dari *website* tanpa harus mencarinya satu persatu.

Pada *search engine* Google, fasilitas ini dinamakan *Google Webmaster Tools*. *Google Webmaster Tools* memfasilitasi agar para *webmaster* dapat memasukkan informasi websitenya seperti *keyword*, sitemap, dan lain sebagainya. Tools ini biasanya digunakan oleh webmaster untuk melakukan optimisasi websitenya agar mudah dikenali search engine atau dikenal dengan metode *Search Engine Optimization* (SEO).

Dewasa ini para *webmaster* berlomba-lomba untuk menerapkan metode *Search Engine Optimization* (SEO) untuk mendapatkan traffic website yang berasal dari *search engine*. Selain itu, dengan menerapkan SEO maka *website* tersebut akan disenangi dan diprioritaskan oleh *Web Crawler* dari search engine untuk dikenali dan dikunjungi dan peluang dari suatu website berada pada indeks teratas dari *search engine* akan semakin besar. Hal ini berbanding lurus dengan kemungkinan bahwa *website* yang berada pada indeks teratas akan semakin sering dikunjungi oleh pengguna *search engine*.

Dalam buku *The Art of SEO: Mastering Search Engine Optimization (Theory in Practice)* ada 11 aspek yang secara umum dapat mempercepat pengenalan *search engine* terhadap suatu website. Aspek-aspek ini dimanfaatkan dalam metode SEO sebagai komponen yang digunakan untuk membuat sebuah website lebih cepat dikenali oleh *search engine*. Aspek-aspek tersebut adalah sebagai berikut :

1. Keunikan alamat atau *url*
2. Kata kunci yang terselip pada *meta tag*
3. Kesesuaian antara kata kunci dengan konten
4. Judul halaman yang unik dan tepat dengan deskripsi yang jelas
5. Kontekstualisasi
6. *Sitemap*
7. *Backlink*
8. Lama loading halaman
9. Pengoptimalan gambar yang ada pada halaman
10. *Update* konten secara berkesinambungan
11. Originalitas konten

Salah satu komponen dari metode SEO diatas adalah *sitemap*. Ada dua jenis *sitemap*, yaitu *sitemap* dengan bentuk *HTML* yang berisi kumpulan *link* pada *website* dan *sitemap* yang berbentuk *XML*. *Sitemap* dengan bentuk *HTML* dapat dengan mudah dikenali oleh pengunjung *website* tersebut sehingga pengunjung dapat mengetahui semua konten yang ada pada *website* tersebut, sedangkan

sitemap dengan bentuk *XML* agak sulit dimengerti oleh pengunjung akan tetapi sangat mudah dikenali oleh *crawler* dari *search engine*.

Pada implementasinya *webmaster* hanya perlu membuat *sitemap* dari *website* yang dimiliki kemudian mengunggahnya pada domain yang sama. Selain itu *webmaster* dapat mendaftarkan *sitemap* websitenya pada *search engine* dan dapat memberikan notifikasi kepada *search engine* tentang lokasi *sitemap* yang telah diunggah sehingga mempermudah search engine mengenali *sitemap* tersebut dan menjadikannya sebagai referensi penelusuran yang dapat mempercepat proses pengenalan sebuah *website* pada *search engine*.

Sitemap dapat dibuat secara manual maupun otomatis. Secara manual dapat dicontohkan bahwa ketika sebuah *website* memiliki indeks berita atau artikel, *webmaster* dapat mengambil *link* dari *database* berita tersebut untuk dibentuk sesuai dengan format *sitemap* yang telah ditentukan oleh sitemaps.org.

Permasalahan yang akan timbul adalah ketika pembuatan sitemap harus dilakukan secara otomatis, dengan *trigger* berupa *url* yang diinputkan oleh user yang bertujuan untuk memudahkan seorang webmaster membuat sebuah sitemap secara instan. Struktur dari setiap *website* mungkin saja berbeda, oleh karena itu dibutuhkan proses tertentu yang dapat mengambil *url* pada halaman *website* tersebut tanpa memperdulikan struktur sebuah *website.*

Dalam pembuatan *sitemap* secara otomatis diperlukan beberapa proses yang dapat mengubah kumpulan *url* yang tersebar pada halaman sebuah *website* menjadi file *sitemap* yang dikenali *search engine*. Proses tersebut adalah alur dari penciptaan *sitemap* yaitu proses *crawling* informasi dengan menggunakan *web*

crawler yang mempunyai fungsi *url recognizer* dan *url grouper* serta proses pembentukan *format sitemap* dari *link* yang telah didapat dari proses *crawling*.

Pada proses *crawling* dengan menggunakan *web crawler* dibutuhkan algoritma tertentu untuk menggerakkan *crawler* menuju *url* halaman *website* untuk mengambil informasi dari halaman tersebut. Menurut M.Najork dalam jurnalnya yang berjudul "*Breadth-First Search Crawling Yields High-Quality Pages*", Algoritma *Breadth-First Search* adalah algoritma alami yang mudah digunakan untuk pencarian informasi pada halaman website terutama dapat membantu proses *crawling* informasi dengan mudah. Penggunaan algoritma ini didasari oleh kemudahan algoritma dan tingkat performansi yang cukup memuaskan.

1.2 Rumusan Masalah

1. Bagaimana implementasi algoritma *Breadth-First Search* dalam pembuatan *Web Crawler*?
2. Bagaimana pengenalan url dilakukan sehingga url dapat dikategorikan sebagai link, backlink, dan filelink?
3. Bagaimana menciptakan *sitemap* yang valid sehingga dapat diterima oleh *search engine*?
4. Bagaimana perbandingan waktu pengenalan mesin pencari Google terhadap sebuah website baru yang menggunakan sitemap dengan yang tidak menggunakan sitemap?

1.3 Batasan Masalah

Untuk memfokuskan penelitian, ditetapkan batasan masalah, sebagai berikut:

1. Pembatasan kedalaman simpul yang dicari ditentukan oleh penulis yaitu 4 tingkat.
2. *Sitemap* yang digunakan menggunakan format *XML*.

1.4 Tujuan Penelitian

Tujuan yang ingin dicapai dalam skripsi ini melakukan implementasi algoritma *Breadth-First Search* pada sistem *Web Crawler* yang bertujuan untuk pembuatan sitemap sebuah *website* dan melakukan implementasi metode *Search Engine Optimization* berbasis sitemap. Adapun detil tujuannya adalah sebagai berikut:

1. Melakukan implementasi algoritma *Breadth-First Search* pada *engine Web Crawler*.
2. Memahami proses pengenalan (recognizing) terhadap url pada sebuah halaman website.
3. Memahami dan melakukan implementasi grouping url menjadi link, backlink dan filelink.
4. Melakukan konversi data hasil *crawling* sesuai format sitemap yang valid sehingga tercipta sebuah alat bantu otomatis pembuatan *sitemap*.
5. Melakukan uji validitas sitemap yang dihasilkan dengan cara mendaftarkan sitemap tersebut ke search engine Google yang telah mengimplementasikan standar sitemaps.org.

6. Mengukur efisiensi waktu pengenalan metode *Search Engine Optimization* dengan menggunakan sitemap terhadap tingkat pengenalan sebuah website baru oleh mesin pencari Google berdasarkan perbandingan lama waktu pendeteksian.

1.5 Manfaat Penelitian

Manfaat yang ingin diperoleh dari skripsi ini adalah:

1. Menyediakan sebuah layanan *Online Sitemap Generator* gratis yang valid dan mudah digunakan, sehingga pengguna dapat menerapkan metode *Search Engine Optimization* menggunakan *sitemap* secara optimal.
2. Menambah wawasan dan dapat menerapkan ilmu yang diperoleh pada waktu kuliah sehingga bermanfaat pada kehidupan nyata.
3. Sebagai bahan referensi bagi para peneliti lain yang ingin mengembangkan *Web Crawler* dan *Sitemap Generator*.

1.6 Metode Penelitian

1.6.1 Alat dan Bahan

Pada penelitian ini digunakan alat penelitian berupa perangkat keras dan perangkat lunak sebagai berikut:

1. Perangkat keras
   a. *Processor* Intel Dual Core 2.20 GHz
   b. RAM 4 GB
   c. *Hard disk* 250 GB
   d. Monitor LED 19”
   e. *Mouse dan keyboard*

2. Perangkat lunak

   a. Notepad ++

   b. XAMPP 1.7.4 (PHP & MySQL)

Bahan penelitian yang digunakan adalah *paper*, *textbook*, dan dokumentasi lainnya yang didapat dari *world wide web*.

1.6.2 Metode Pengumpulan Data

Untuk memperkuat Web Crawler beserta fungsi maka diperlukan metode penelitian pengumpulan data sebagai referensi dengan metode studi kepustakaan, yaitu dengan mempelajari literatur berupa artikel, *paper* maupun sumber lain yang berhubungan dengan objek penelitian.

1.6.3 Metode pengembangan Perangkat Lunak

Metode pengembangan perangkat lunak yang digunakan untuk membangun perangkat lunak Sitemap Generator ini adalah *Sequential Linear* (Pressman, 2001. h.28).

1.7 Sistematika Penulisan

Sistematika penulisan proposal ini disusun untuk memberikan gambaran umum tentang perangkat lunak yang akan dibuat. Sistematika penulisan tugas akhir ini adalah sebagai berikut :

BAB I PENDAHULUAN

Bab ini menguraikan tentang latar belakang masalah, rumusan masalah, maksud dan tujuan, batasan masalah, metode penelitian dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Bab ini menguraikan beberapa hal yaitu landasan teori berupa pengertian *search engine*, *web crawler*, *sitemap*, basis data, perancangan sistem, pengertian internet, pengertian *world wide web*, pengertian HTTP, bahasa pemrograman yang digunakan, algoritma *Breadth-First Search*, tinjauan perangkat lunak dan lain sebagainya.

BAB III METODE PENELITIAN

Bab ini memaparkan tentang rancangan penelitian, fokus penelitian, dan metode yang digunakan dalam penelitian.

BAB IV IMPLEMENTASI

Pada bab ini diuraikan tentang lingkungan implementasi, implementasi antar muka, pengujian perangkat lunak yang menggunakan pengujian *Black Box*.

BAB V KESIMPULAN DAN SARAN

Pada bab ini berisi tentang kesimpulan dari penelitian implementasi web crawler dengan algoritma *Breadth-First Sear*ch sebagai *sitemap generator* dan saran yang diajukan penulis agar dapat menjadi bahan pertimbangan peneliti lain.