BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Search engine atau mesin pencari adalah sebuah program yang dibuat untuk melakukan pencarian pada situs web. *Search engine* pertama kali diperkenalkan oleh Alan Emtage pada tahun 1990 dengan nama *Archie*. Prinsip keja *Archie* adalah melakukan pengindeksan semua *file* pada web. Lokasi file yang dicari akan lebih mudah ditemukan dengan *Archie* oleh para pengguna internet (Seymour et al., 2011).

Archie dapat menampilkan daftar nama situs, tetapi tidak dapat menampilkan isi atau konten. Pada tahun berikutnya mulai bermunculan *search engine* baru. *Aliweb* muncul pada tahun 1993, para pengguna diberi kesempatan untuk mengunggah halaman situs yang ingin terindeks di internet, dan dapat mengisi deskripsi untuk halaman tersebut. Kemudian, muncul *AltaVista* yang populer di internet pada tahun 1995. *AltaVista* memberikan *unlimited bandwith* untuk penggunanya, teknik dan sistem algoritmanya juga sudah lebih maju. Sembilan tahun setelahnya, muncul *Yahoo*. Data dari situs-situs yang sudah dikenal banyak orang tersedia di *Yahoo*. Pengelola situs dan pemilik situs dapat menambah informasi tanpa mengeluarkan biaya, terdapat juga fitur yang sudah lengkap dengan dikenakan sejumlah biaya. Perusahaan Yahoo sangat lambat dalam pengembangan mesin pencarian dan para pengguna internet perlahan mulai mencari *search engine* lain (Seymour et al., 2011).

Search engine yang saat ini masih populer yaitu *Google*. *Google* diluncurkan pada tahun 1998, *Google* mengembangkan sistem yang berbeda dengan *search engine* sebelumnya. Sistem yang dikembangkan bernama *BackRub*, sistem ini

menggunakan *backlink* untuk memberi peringkat pada setiap halaman. *Ranking* untuk setiap halaman situs diurutkan berdasarkan penyebutan situs terbanyak di situs-situs lain (Brin et al., 1998). Selain *Google*, *search engine* yang masih populer saat ini yaitu *Bing*. *Bing* muncul di pertengahan 2009. *Bing* merupakan salah satu *search engine* favorit di Amerika Serikat, karena terdapat beberapa fitur yang pas digunakan masyarakat Amerika. Namun, selain di Amerika, *Google* lebih populer dan sangat favorit dibandingkan dengan *Bing* (Seymour et al., 2011).

Tabel 1.1: Penggunaan *global search engine* Agustus 2020 (NetMarketShare, 2020)

Search engine	**Global share (%)**
Google	83.46
Baidu	7.35
Bing	6.15
Yahoo!	1.42
Yandex	0.87
DuckDuckGo	0.31

Apple adalah perusahaan yang saat ini menggunakan *Google* sebagai mesin pencarinya. Selama beberapa tahun ini *Google* membayar *Apple* lebih dari satu miliar dolar untuk tetap menjadi *search engine* utama di *safari* untuk *iOS*, *iPadOS*, dan *macOS*. *Apple* sudah membuat *web crawler* bernama *Applebot* untuk *search engine Apple* dan kemungkinan *search engine* tersebut akan rilis di tahun ini (Henshaw, 2020).

Fungsi dari *search engine* selain mencari informasi adalah sebagai tempat untuk memasangkan iklan. Karena jumlah pengguna yang terus bertambah, menjadikan *search engine* sebagai media pemasaran saat ini. Melalui *search engine*, pengguna internet dapat mencari apa saja, termasuk juga barang yang mereka ingin

beli. *Search engine* membuat pencarian informasi semakin mudah. Banyaknya pengguna dapat menjadikan *search engine* sebagai peluang bagi pemilik usaha atau perusahaan untuk memasarkan produk dan jasa mereka.

Google memasangkan iklan pada mesin pencarinya untuk mendapatkan keuntungan, layanan iklan ini bernama *AdWords*. *AdWords* memungkinkan pengiklan menjangkau pengguna *online* dengan beriklan di platform *Google*. Jika mencari sebuah informasi suatu barang, maka beberapa baris hasil penelusuran teratas yang disematkan dengan tulisan "Ad" akan terlihat. Informasi tersebut dipasang oleh pengiklan yang menggunakan fitur *AdWords Google*.

Google pun juga mendapatkan banyak data melalui mesin pencariannya. Selain itu, *Google* juga membuat layanan lain seperti *Gmail*, *YouTube*, *Search*, *Drive*, *Maps*, dan *PlayStore*. Setiap layanan memiliki fungsi tersendiri untuk membantu aktivitas masyarakat. Dan dari setiap layanan *Google* terdapat data-data pengguna yang dapat dipakai untuk keperluan lainnya.

Proses pembuatan arsitektur *search engine* ini dimulai dengan membuat salah satu bagiannya, yaitu *crawler*. Sebelum akhirnya *search engine* dapat menampilkan data-data yang diperlukan *user*, *search engine* memerlukan sebuah bagian yang berfungsi untuk mengumpulkan data-data tersebut. Maka dibutuhkan sebuah *crawler* yang akan dirancang dengan sebaik mungkin.

Web crawler merupakan *software* yang bekerja otomatis untuk menjelajahi *World Wide Web* secara terorganisir. Web crawler pertama muncul pada tahun 1944 bernama *RBSE* (Eichmann, 1994). Web crawler ini didasarkan pada dua program, yaitu:

1. *Spider* yang berfungsi memelihara *queue* dalam *relational database*.
2. *Mite* untuk mendownload halaman dari web.

Fungsi dari *web crawler* adalah untuk mengumpulkan data. Jika tidak ada *web crawler*, maka *search engine* tidak akan mendapatkan data dan melakukan *indexing*. Data-data yang dihasilkan oleh *web crawler* merupakan data terbaru dan akurat. *Google* memiliki *web crawler* bernama *Googlebot*. Selain *Googlebot*, terdapat *web crawler* lain yaitu: *Bingbot* (*web crawler Bing*), *Slurp bot* (*web crawler Yahoo!*), *DuckDuckBot* (*web crawler DuckDuckGo*, *Baiduspider* (*web crawler Baidu*), *Yandex Bot* (*web crawler Yandex*).

Kaur dan Geetha (2020) dalam jurnalnya, telah membuat *web crawler* untuk *hidden web* dengan menggunakan *SIM+HASH* dan *Redis Server*. *Focused web crawler* dapat memberikan hasil halaman relevan yang maksimum. Akan tetapi, *crawler* ini sudah terlalu rumit untuk pembuatannya. Komponen pada *focused web crawler* hampir sama dengan *web crawler* pada umumnya.

Skripsi ini dikerjakan bersama dengan Savira Rahmayanti dari Ilmu Komputer angkatan 2015 Universitas Negeri Jakarta, akan mengembangkan *server based search engine* yang bersifat *open source* dan menjelaskan seperti apa arsitektur *search engine* tersebut. Penulis mendapatkan bagian untuk mengimplementasi *web crawling* pada *search engine*, sedangkan Savira fokus untuk mengerjakan bagian *searching* dan *indexing*.

Pada skripsi ini tidak mengimplementasikan apa yang telah dikembangkan oleh Sawroop Kaur dan Geetha (2020), karena *web crawler* tersebut sudah terlalu rumit dan komponennya tidak jauh berbeda dengan *crawler* biasa. Akan tetapi, skripsi ini akan mengembangkan *web crawler* dasar yang merujuk pada awal perkembangan *search engine Google* (Brin and Page, 1998). Penelitian ini akan menjadi dasar pada penelitian *search engine* berikutnya.

Metode algoritma *crawler* pada penelitian ini, akan mengacu pada awal perkembangan *Google*, yaitu *modified similarity based crawling* (Cho et al., 1998).

Karena algoritma ini sangat pas untuk merancang sebuah *crawler* yang membangun database dengan topik tertentu. Sebagai contoh, jika ingin membangun *crawler* dengan topik *Barcelona*, maka *crawler* akan *crawling* halaman yang terdapat kata Barcelona terlebih dahulu. Algoritma *modified similarity based crawling* sangat menarik untuk diimplementasikan pada penelitian *search engine* ini.

Crawler sangat dibutuhkan *search engine* untuk dapat mencari informasi dengan lebih efisien. Skripsi ini juga untuk memperdalam ilmu mengenai *information retrieval* khususnya *web crawling*. Oleh karena itu, perlu dibuat perancangan *web crawler* pada *search engine*. Dan ini tertuang pada penelitian yang berjudul **"Perancangan *Crawler* Sebagai Pendukung pada *Search Engine*"**.

1.2 Rumusan Masalah

Berdasarkan uraian pada latar belakang yang diutarakan di atas, maka perumusan masalah pada penelitian ini adalah "Bagaimana cara membuat rancangan *crawler* sebagai pendukung pada *search engine*?"

1.3 Pembatasan Masalah

Adapun batasan-batasan masalah yang digunakan agar lebih terarah dan sesuai dengan yang diharapkan serta terorganisasi dengan baik adalah:

1. Penelitian ini hanya membuat sebagian arsitektur *search engine*, yaitu *crawler*. Algoritma *crawler* yang akan dibuat mengacu pada model algoritma *Google* awal.
2. *Crawler* tidak mengambil data dari isi database web, melainkan dari struktur data *HTML* web. Maka *crawler* hanya akan berjalan pada *website statis*.

3. Penelitian ini akan menargetkan dua situs. Situs pertama menggunakan *HTML* versi 4, dan situs kedua menggunakan *HTML* versi 5.

1.4 Tujuan Penelitian

1. Membuat *crawler* yang dipakai untuk kebutuhan *search engine*.
2. Untuk mengetahui arsitektur *search engine*.
3. Untuk mempelajari cara kerja *crawling*.

1.5 Manfaat Penelitian

1. Bagi penulis

   Menambah pengetahuan dibidang *information retrieval* khususnya mengenai *search engine* dan *crawling*, mengasah kemampuan *programming*, dan memperoleh gelar sarjana dibidang Ilmu Komputer. Selain itu, penulisan ini juga merupakan media bagi penulis untuk mengaplikasikan ilmu yang didapat di kampus ke kehidupan masyarakat.

2. Bagi Program Studi Ilmu Komputer

   Penelitian ini menjadi langkah awal penelitian *search engine* berikutnya, dan dapat memberikan gambaran bagi seluruh mahasiswa khususnya bagi mahasiswa program studi Ilmu Komputer Universitas Negeri Jakarta tentang proses perancangan *crawler*.

3. Bagi Universitas Negeri Jakarta

   Menjadi pertimbangan dan evaluasi akademik khususnya Program Studi Ilmu Komputer dalam penyusunan skripsi sehingga dapat meningkatkan kualitas

akademik di program studi Ilmu Komputer Universitas Negeri Jakarta serta meningkatkan kualitas lulusannya.