BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Penelitian tentang teknik Crawling pernah dilakukan oleh Christopher Olston dan Marc Najork (2010) di University of California, Barkeley. Kedua peneliti tersebut membahas mengenai survey dan penelitian ilmiah tentang pembelajaran Web crawling dimana merujuk pada web yang menggunakan *Breadth-first search* atau aplikasi yang menggunakan algoritma yang melakukan pencarian secara meluas yang mengunjungi situs secara *preorder* dengan menentukan antrian yang harus dikerjakan terlebih dahulu dengan tantangan web crawl data yang besar dengan implementasi *state-of-the-art* atau pencapaian paling tinggi dari sebuah proses pengembangan.

Penelitian kedua dilakukan oleh Dewi Rosmala dan Rizqi Rivani Syafei (2011) di institut Teknologi Nasional Bandung. Penelitian ini membahas tentang proses menghimpun, memahami dan merespon opini tentang merk,produk,reputasi atau opini user di sosisal media dengan tujuan menjaga brand image produk dengan menganalisis menggunakan web crawler untuk mencari aktifitas dan pembicaraan yang sedang terjadi dan menyelesaikan dengan mencari cara untuk mempengaruhi dan membentuk opini di sosial media.

Penelitian ketiga dilakukan oleh Andrei Z. Broder (2003) di IBM TJ Watson Research Center, Hawthorne, New York. Penelitian ini membahas mengenai dasar penerapan web crawling pada proses penangkapan saat parsing data dari url target dengan menganalisis jumlah halaman web dan prosentase

perubahan halaman target per minggu dengan melakukan riset mengenai manfaat caching dengan menggunakan teknik url caching untuk web crawling.

Penelitian keempat Carlos Castillo tahun (2004) di University of Chile. Secara garis besar penelitian ini membahas mengenai bagaimana pengimplementasian Web Crawling dari sudut pandang keefektifan web crawling dalam pengalokasian informasi dan menampilkan informasi serta prosentase web yang dapat ter-*crawl* berdasarkan domain, seperti .cl .gr domain untuk Negara Chili dan Greece ataupun .edu ataupun .com mengenai informasi pengaksesan domain, dimana analisis yang diambil berdasarkan seberapa efektif data yang diperoleh mulai dari isi data yang ter-enskripsi dan penerapan network bandwidth yang begitu besar untuk mengunduh suatu halaman yang diakses dan mempertimbangkan penyajian informasi dengan Bandwidth yang lebih ringan atau kecil.

Penelitian kelima dilakukan oleh Nurhayati Masthurah, Taufiq Wirahman, Devi Munandar (2013) di Pusat Penelitian Informatika Lembaga Ilmu Pengetahuan Indonesia. Penelitian ini memanfaatkan Website Parser Template (WPT) yang terdiri dari berbagai bagian melipui ontologi,template dan url dimana ontologi mencakup semua konsep dan berhubungan dengan keseluruhan yang ada di dalam web dengan mengumpulkan data dengan menggunakan sistem pencarian berbasis semantik dengan metadata RDF (Resource Description Framework) yaitu framework yang mendefinisikan resource di dalam web. kumpulan RDF ini lah yang digunakan sebagai Repositori data untuk membangun Semantic web.

Dari ke lima referensi diatas, ditemukan perbedaan antara penerapan web crawling yang telah dibuat dengan yang akan dibuat yaitu implementasi yang dikhususkan untuk menyaring informasi atau data dari situs jual beli online yang nantinya akan disajikan didalam sebuah website yang dapat memudahkan user atau konsumen jual beli online dalam memilih Smartphone sesuai kebutuhan harga yang diperlukan.

Perbandingan dengan penelitian sebelumnya, yang terkait dengan penelitian yang akan dilakukan dapat dilihat pada Tabel 2.1.

Tabel 2.1 Perbandingan dengan penelitian sebelumnya

No	Peneliti	Objek	Tujuan Penelitian	Teknik yang digunakan	Informasi Yang dihasilkan
1	Christopher Olston, Marc Najork	"Breadth-first search" Web	Pengembangan implementasi *state-of-the-art* web crawling	Teknik url caching	Hasil survey mengenai jumlah pengaksesan konten dari berbagai web
2	Dewi Rosmala, Rizqi Rivani Syafei	Media Sosial Twitter	Menjaga *Brand Image Product*	Pemanfaatan akses API (Application programming interface)	Hasil pemantauan mengenai isu-isu brand image dari sebuah produk
3	Andrei. Z Broder	World Wide Web Cahce	Mengetahui Pemanfaatan Web caching	Teknik url caching	Prosentase perubahan halaman target per minggu
4	Carlos Castillo	Crawling Web berdasarkan Web domain	Keefektifan web crawling dalam pengalokasian informasi dan menampilkan informas	Analisis network bandwidth	Informasi dan prosentase web yang dapat ter-crawl berdasarkan domain

Tabel lanjutan Tabel 2.1

5	Nurhayati Masthurah, Taufiq Wirahman	Website Parser Template (WPT)	Pengumpulan Metadata sebagai Repositori data untuk membangun Semantic web.	Teknik Scrapping	Kumpulan RDF dalam Semantic web
6	Yang disulkan	Konten Penyedia Layanan jual beli online	Pengimplement asian dasar web crawling	Teknik Scrapping	Website Daftar Harga Smartphone dari berbagai situs jual beli online

2.2 Dasar Teori

2.2.1 Web Crawler

Web crawler adalah sebuah perangkat lunak yang digunakan untuk menjelajah serta mengumpulkan halaman-halaman web yang selanjutnya diindeks oleh mesin pencari (Gatial E, 2005). Sering juga disebut dengan web spider atau web robot yang menjelajahi halaman website dengan menggunakan web browser.

Sebuah program yang memberikan suatu atau lebih *seed url* yang akan mengunduh halaman-halaman dari url yang dituju dan mengekstrak beberapa konten yang ada didalamnya (Najork M, 2010).

Menurut Dipanegara Computer Club yang dikutip oleh Muhammad ikhsan (2011) *Web Crawler* lebih dikenal sebagai sebuah program/script otomatis yang memprosess halaman web. Bisa juga disebut sebagai web spider atau web robot, berfungsi mengidentifikasi hyperlink dan melakukan proses kunjungan/visit secara rekursif.

2.2.2 Parsing

Parsing atau sintaksis adalah sebuah mesin yang akan menyaring data yang bersifat meta-bahasa untuk mengurai item yang akan dicari dalam pencarian berbasis komputer dengan aplikasi tertentu kedalam situs web dan diakses menggunakan jaringan internet (Bernard A. Ferret. 2002).

Menurut Charniak dan Johnson (2005) Parsing adalah sebuah penguraian struktur sintaksis dari sebuah string, membeberkan data yang telah terlabel dengan berbagai macam teknik yang digunakan.

Parsing adalah suatu teknik untuk memisahkan suatu teks dari tag kode dalam html pada halaman website .atau juga yang biasa di sebut Screen Scraper yaitu teknik untuk mengambil isi sebuah halaman web secara spesifik.
Dipanegara Computer Club. 2011. Python | Web Scraping & Parsing/Screen Scraping Web Pages.

2.2.3 Web Scraping

Web Scraping (Turland, 2010) adalah proses pengambilan sebuah dokumen semi-terstruktur dari internet, umumnya berupa halaman-halaman web dalam bahasa markup seperti HTML atau XHTML dan menganalisis dokumen tersebut untuk untuk diambiol data tertentu dari halaman tersebut untuk digunakan bagi kepentingan lain.

Web Scraping sering dikenal dengan *screen scraping*. *Web Scraping* tidak dapat dimasukkan dalam bidang data mining, karena data mining menyiratkan upaya untuk memahami pola semantik atau tren dari sejumlah besar data yang

telah diperoleh yang berfokus pada cara memperoleh data melalui pengambilan dan ekstraksi data dengan ukuran data yang bervariasi (Jot et al, 2012).

2.2.4 PHP

Pretext Hyper-Processor (Antonius Nugraha Widhi Pratama.2010) PHP adalah adalah bahasa *scripting* yang menyatu dengan HTML dan dijalankan pada *server side*. Artinya semua sintaks yang kita berikan akan sepenuhnya dijalankan pada server sedangkan yang dikirimkan ke browser hanya hasilnyasaja. PHP menyatu dengan bahasa HTML untuk membuat halaman web yang menarik (Antonius Nugraha Widhi Pratama, 2010).

2.2.5 CURL

CURL merupakan librari php yang memungkinkan untuk mentransfer data melalui berbagai protocol dan banyak digunakan sebagai cara untuk mengirim atau meminta data dari satu atau beberapa situs, permintaan dengan *CURL* tidak dibatasi dalam hal apapun, mirip seperti *HTTP* dasar dan dapat mengupload *FTP* serta memungkinkan untuk melakukan atifitas yang lebih lebih kompleks seperti interaksi otentifikasi dengan situs *HTTPS* tertutup. (Sojish Krishnan, 2006)

Teknik *CURL* hampir sama dengan mengirim data menggunakan *GET method* yaitu menggunakan *URL*, Secara umum, penggunaan *CURL* untuk mengirim beberapa data dengan *POST method* ke server harus menentukan konfigurasi yang diperlukan untuk setiap *CURL*. (Rachmad Andri Atmoko.2005)

2.2.6 DOM

Document Object Model (DOM) merupakan sebuah ketentuan yang dikembangkan oleh *W3C* untuk berinteraksi dengan objek-objek yang ada di dalam *HTML, XML,* maupun *XHTML*. *DOM* bersifat *cross-platform* dan *language-independent,* artinya *DOM* dapat digunakan dengan bahasa pemrograman apapun, dalam sistem operasi manapun. (Alex Xandra Albert Sim.2014)

2.2.7 Simple HTML DOM

PHP simple HTML DOM merupakan sebuah wadah bagi pengembang *php* maupun *DOM* sebagai *parser* data dengan kegunaan memudahkan pengambang dalam setiap aktifitas *parsing* data melalui pemrograman *php* dengan pencarian struktur atau elemen *DOM* yang dapat dengan mudah di identifikasikan dan diterjemahkan dalam pemrograman menggunakan *php*. (David walsh.2011).

Menurut Paulus Setyo (2015) Simple html Dom adalah sebuah metode untuk melakukan parsing html data dengan beberapa skenario yang membutuhkan penerjemahan data dalam bahasa pemrograman php dengan tujuan pengambilan data yang memungkinkan pengubahan struktur penulisan dan diterjemahkan dalam bahasa pemrograman php.