BAB 2

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Penelitian dilakukan oleh Masthurah dan Munandar (2013) di Pusat Penelitian Informatika Lembaga Ilmu Pengetahuan Indonesia, Penelitian ini memanfaatkan *Website Parser Template (WPT)* template dan *url* dimana teknologi 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.

Penelitian kedua tentang *teknik Crawling* pernah dilakukan oleh Olston dan 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 s*tate-of-the-art* atau pencapaian paling tinggi dari sebuah proses pengembangan.

Penelitian ketiga dilakukan oleh 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 dilakukan oleh Rosmala dan 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 kelima dilakukan oleh Triawan (2016) Penelitian ini membahas mengenai mengimplementasikan teknik *Crawling* secara terstruktur serta memberikan langkah-langkah teknik *Crawling* tanpa menggunakan akses *API* (*Application programming interface*) yaitu melalui parsing data dengan teknik *Scraping* untuk melalukan riset mengenai perbandingan harga *smartphone* yang diharapkan bisa mempermudah konsumen dalam menentukan harga produk yang dicari hanya dalam suatu wadah berupa website dengan menampilkan informasi produk yang akan dikomparasikan dari beberapa situs jual beli online.

Dari ke lima referensi diatas, ditemukan perbedaan antara penerapan web crawling atau web scraping yang telah dibuat dengan yang akan dibuat yaitu belum ada yang menerapkan metode web *scraping* untuk barang promo pada toko online di indonesia.

Penelitian yang dilakukan sekarang bisa dilihat pada table 2.1.

Tabel 2.1 Tabel Perbandingan Penelitian

No	Peneliti	Objek	Tujuan Penelitian	Teknik yang digunakan	Informasi Yang dihasilkan
1	Masthurah dkk (2013)	*Website Parser Template (WPT)*	Pengumpulan Metadata membangun *Semantic web*	*Teknik Scrapping*	Kumpulan *RDF* dalam *Semantic web*
2	Olston dan Najork (2009)	"*Breadth-first search" Web*	Pengembangan implementasi *state-of-the-art web crawling*	*Teknik url caching*	Hasil survey mengenai jumlah pengaksesan konten web
3	Broder (2003)	*World Wide Web Cahce*	Mengetahui Pemanfaatan *Web caching*	*Teknik url caching*	Prosentase perubahan halaman target per minggu
4	Rosmala, dan Rivani (2011)	Media Sosial Twitter	Menjaga Brand Image Product	Pemanfaatan akses *API*	Hasil pemantauan brand image dari sebuah produk
5	Triawan (2016)	Konten Layanan jual beli online	Pengimplement asian dasar *web crawling*	*Teknik Web Scrapping*	Website Daftar HargaHP dari berbagai jual beli online
6	Francisco (2019)	Penyedia Promo toko online	Pengimplement asian dasar *web srcaping*	*Teknik Web Scrapping*	Website Daftar promo pada 10 toko online

2.2 **Dasar Teori**

2.2.1 **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 diambil 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.

2.2.2 **CodeIgniter**

CodeIgniter merupakan aplikasi *open source* berupa *framework PHP* dengan model *MVC (Model, View, Controller)* untuk membangun aplikasi web dinamis dengan cepat dan mudah, *CodeIgniter* memiliki desain dan struktur file yang sederhana, didukung dengan dokumentasi yang lengkap sehingga *framework* ini lebih mudah di pelajari.

CodeIgniter ini memungkinkan para pengembang untuk menggunakan *framework* secara parsial atau secara keseluruhan, artinya bahwa *CodeIgniter* masih memberi kebebasan kepada para pengembang untuk menulis bagian-bagian kode tertentu di dalam aplikasi menggunakan cara konvensional atau dengan

syntax umum didalam *PHP,* tidak harus menggunakan aturan penulisan kode di *CodeIgniter,* (Septian, 2011).

2.2.3 **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, 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* atau *Screen Scraping Web Pages.*

2.2.4 ***PHP***

Pretext Hyper-Processor 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 hasilnya saja, *PHP* menyatu dengan bahasa *HTML* untuk membuat halaman web yang menarik.

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, (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,* (Atmoko, 2005).

2.2.6 ***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.* (Walsh, 2011).

Menurut Setyo, (2015) *Simple html Dom* adalah sebuah metode untuk melakukan *parsing html* data dengan beberapa skenario yang membutuhkan penerjemahan data dalam bahasa pemrograman *php d*engan tujuan pengambilan

data yang memungkinkan pengubahan struktur penulisan dan diterjemahkan dalam bahasa pemrograman *php.*