Perbandingan Metode Web Scraping Menggunakan CSS Selector dan Xpath Selector

Taufiq Rizaldi
Jurusan Teknologi Informasi
Politeknik Negeri Jember
taufiq_r@polije.ac.id

Hermawan Arief
Jurusan Teknologi Informasi
Politeknik Negeri Jember
hermawan_arief_putranto@yahoo.com

Abstrak - Pemanfaatan data atau berita yang tersebar di internet untuk meningkatkan peluang keberhasilan dalam sebuah usaha melalui analisa trend pasar adalah hal yang sangat umum pada saat ini. Penjelajahan Web (*Crawl*) dan ekstraksi data dari web (*Scraping*) menjadi salah satu hal yang penting, agar tidak terjadi data yang kurang sempurna, dan data yang diterima adalah data yang paling baru. *CSS Selector* dan Xpath merupakan salah satu metode yang umum digunakan dalam melakukan proses *crawling*. Terdapat perbedaan dari jumlah data yang terambil, besar *file* output dan waktu pemrosesan dari kedua metode tersebut, dimana Xpath memiliki keunggulan pada jumlah data yang terambil dan waktu pemrosesnya yang berakibat pada ukuran file output yang lebih besar. Sedangkan untuk penggunaan memori pada kedua metode pada proses *crawling* tidak memiliki perbedaan yang signifikan

Kata Kunci : *Web Crawling, Web scraping, Scrapy, Xpath, CSS Selector.*

I. PENDAHULUAN

Internet adalah sebuah tempat berkumpulnya sejumlah besar informasi di dunia, baik itu teks, media atau data dalam format lain yang biasanya ditampilkan dalam sebuah halaman web. Kemudahan untuk mengakses data tersebut sangat penting bagi keberhasilan sebagian besar bisnis di dunia modern. Bagi perusahaan yang bergerak dibidang pemasaran, data tersebut bisa digunakan untuk mengetahui trend pasar yang berkembang saat ini, sehingga bisa diketahui strategi pemasaran yang paling tepat untuk tiap produk. Bagi perusahaan yang berbasis *E-commerce* juga bisa memanfaatkan data tersebut untuk analisis pasar atau sekedar perbandingan harga dengan kompetitor ecommecre lain.

Keberadaan data dalam jumlah besar dan beragam juga mendorong beberapa peneliti untuk menggali informasi yang tersirat atau melakukan proses analisis pada fenomena tersebut. Diantaranya adalah penelitian tentang analisis sentiment berbasis ontology untuk mengukur persepsi produk yang menggunakan *microblogging tweeter* sebagai sumber datanya [1]. Pada penelitian tersebut, data yang berupa kicauan *(tweet*) diperoleh melalui layanan yang disediakan oleh *microblogging tweeter* yang disebut *Tweeter API*. Namun tidak semua website yang ada di internet memberikan layanan tersebut, dan itu menjadi masalah tersendiri apabila ada yang ingin menakses data mereka. Berikutnya adalah penelitian tentang klasifikasi dokumen berita yang memuat konten *E-Government* menggunkan metode *Naïve Bayes Classifier* [2]. Dalam penelitian tersebut, digunakan portal berita nasional www.jawapos.com sebagai sumber data yang digunakan untuk proses *training*.

Sayangnya sebagian besar data di internet memiliki hak akses yang sangat terbatas. Tidak seperti *microblogging tweeter* yang memiliki *Tweeter API*, sebagian besar situs web di internet tidak menawarkan opsi untuk menyimpan data yang mereka tampilkan ke penyimpanan lokal komputer, atau ke situs web pribadi. Untuk mengakses data dari situs-situs seperti itu dibutuhkan teknik khusus yaitu *scraping.*

Web scraping merupakan teknik yang digunakan untuk mengekstrak sejumlah besar data dari situs web dimana data yang sudah diekstraksi disimpan ke sebuah file lokal di komputer atau ke database dalam format tabel (*spreadsheet*). Inilah yang memungkinkan user untuk mengeksplorasi isi dari situs web tanpa mengunjungi situs web yang bersangkutan, sehingga *user* bisa melakukan berbagai bentuk analisis tanpa mengganggu *resource* situs web yang bersangkutan.

Banyak *tool* yang bisa digunakan untuk melakukan web scraping, salah satu yang populer adalah scrapy. Scrapy adalah sebuah *framework* aplikasi yang digunakan untuk menjelajahi (*crawling*) situs web dan mengekstrak data terstruktur sehingga dapat digunakan untuk berbagai aplikasi lain yang bermanfaat, seperti *data mining*, pemrosesan informasi atau arsip sejarah [3]. Walaupun bersifat open source, namun scrapy merupakan *framework web scraping* yang handal dan fleksibel, sehingga hanya dibutuhkan sedikit penyesuaian apabila kita ingin menjelajahi beberapa situs yang berbeda. Hal yang biasa dilakukan pada saat melakukan *web scraping* adalah mengekstraksi data dari halaman web yang berbentuk dokumen html. Scrapy memiliki mekanisme tersendiri untuk mengekstrak data dari dokumen html yang disebut selector karena mereka "memilih" bagian tertentu dari dokumen HTML yang ditentukan baik oleh ekspresi XPath maupun CSS. XPath adalah bahasa untuk memilih simpul (node) dalam dokumen XML, yang juga bisa digunakan dengan HTML. CSS adalah bahasa untuk menerapkan style pada dokumen HTML. Dalam paper ini akan disajikan perbandingan hasil dari penggunaan kedua mekanisme tersebut, sehingga dapat diketahui manakah yang paling baik

digunakan, apabila kita ingin mengekstrak data dari sebuah situs yang tidak menyediakan layanan ekstraksi data.

Dari uraian diatas, permasalahan yang diangkat dalam penelitian ini adalah bagaimana membuat mekanisme yang dapat membandingkan penggunaan XPATH dan CSS *selector* sebagai metode web *scraping* menggunakan scrapy. Hal ini membawa kedalam permasalahan yang lebih rinci, yang pertama adalah, bagaimana mengimplementasikan Scrapy sebagai *web scraper* yang akan menjelajah situs tertentu. Kedua, bagaimana membuat *spider* yang mengimplementasikan kedua metode seleksi tersebut. Yang ketiga adalah bagaimana menyajikan data yang dihasilkan dari penerapan metode XPATH dan CSS *Selector* sehingga diketahui metode mana yang paling tepat untuk digunakan.

II. METODOLOGI PENELITIAN

A. *Web Crawler*

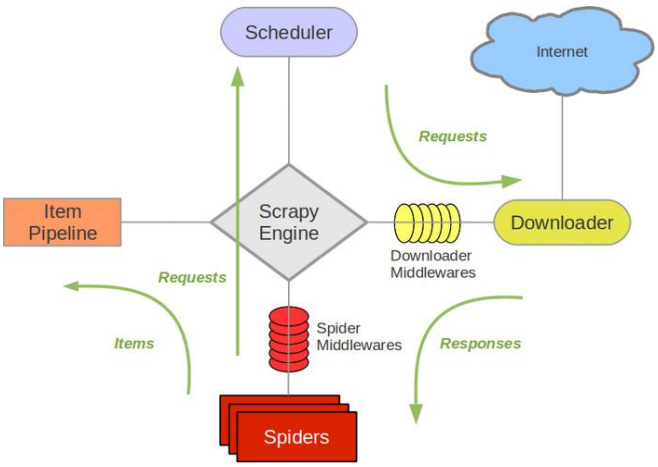
Web Crawler adalah suatu program atau *script* otomatis yang relatif sederhana, yang dengan mengunakan metode tertentu melakukan *scan* ke semua halaman-halaman Internet untuk membuat indeks dari data yang dicarinya. Pada umunya *crawling* diterapakn pada web yang banyak disebut *Web Crawling. Web Crawling* pada umumnya digunakan pada search engine yang dilakukan oleh sekelompok kompuer yang dikluster dimana setiap komputer menjalankan beberapa *thread* [4].

Prosedur dari *Web Crawling* dimulai dari memilih satu set URL yang akan dilakukan proses *crawling* dimana URL tersebut menyediakan banyak link ke halaman yang penting untuk proses *Crawling. Crawler* akan mengunduh konten dari halaman tersebut ke dalam penyimpanan lokal seperti *hardisk* dll. Secara bersaman *thread* yang ada pada c*rawler* mencari link baru dari halaman yang terhubung dengan halaman yang telah diunduh, *link* baru membentuk yang disebut dengan *crawler's frontier*. Tujuan utama dari *crawler's frontier* adalah mendapatkan halaman baru sebanyak mungkin dan *update* tingkat kebaruan halaman yag telah diunduh.

B. Scrappy

Scrapy adalah sebuah *framework* yang digunakan untuk melakukan proses *crawling* dan mengekstrak data yang tersruktur. Scrapy digunakan pada proses *data mining*, pemrosesan informasi dan pengarsipan *history*. Scrapy dibangun dengan menggunakan Pyhton yang di *support* dengan twisted [5]. Terdapat tujuh komponen utama pada scrappy seperti yang ditunjukan pada gambar 1, yaitu *Scheduler, Item Pipeline, Downloader, Downloader Middleware, Spiders, Spiders Middleware.*

Scrapy *Engine* bertanggung jawab untuk mengendalikan arus data antar semua komponen sistem. *Downloader* bertangung jawab untuk mengambil halaman web yang diminta dan memasukananya ke dalam *engine*. *Spiders* adalah sebuah *class* yang dibuat oleh *user* untuk memindah respon yang didapat dari *engine* dan mengektrak item dari respon tersebut. *Pipeline* bertanggung jawab untuk memproses *item* setelah *item* tersebut terekstrak oleh *spiders*. *Downloader middlewares* adalah perantara atau jembatan yang berada diantara *engine* dan *downloader* yang bertugas memproses *request* dari *engine* ke *downloader* dan memberikan respon dari *downloader* ke *engine*. *Downloader middlewares* menyediakan mekanisme yang sesuai untuk memperluas fungsi Scrapy dengan memasukkan kode yang dapat diubah sesuai dengan kebutuhan.



Gambar 1. Tujuh Komponen Utama Scrapy

III. HASIL DAN PEMBAHASAN

Sebelum dilakukan implementasi program, perlu dilakukan analisa dan desain sistem untuk mempermudah implementasi program karena sebagai acuan untuk menghasilkan program yang baik.

A. Objek Penelitian

Dalam penelitian ini, objek penelitiannya adalah sebuah *weblog* yang bernama blogdetik (http://blog.detik.com/). Blog ini disediakan oleh salah satu situs berita populer di Indonesia detik.com (http://www.detik.com) sebagai wadah untuk menampung karya tulis seluruh *blogger* di Indonesia, baik yang sudah memiliki *blog* sendiri maupun yang belum.

Ada tiga kategori dalam *blog* ini yang digunakan sebagai objek *scraping* yaitu, komunitas, hiburan dan kuliner. Hanya artikel yang berada dibawah kategori tersebut yang akan diekstrak. Hal ini untuk mengetahui apakah scrapy juga bisa digunakan sebagai *web scraping* terbimbing.

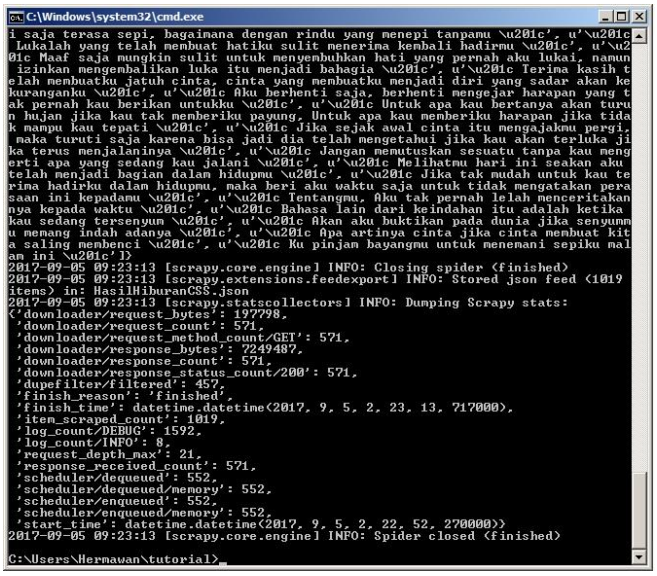
B. Variabel Penelitian

Dari semua artikel yang berada didalam blogdetik, hanya data tertentu yang akan diekstrak dan disimpan kedalam beberapa variabel. Variabel yang pertama adalah *link* yang berisi tautan untuk menuju halaman web yang memuat artikel berita. Yang kedua adalah judul, yang berisi judul artikel, kemudian berikutnya adalah deskripsi yang berisi deskripsi singkat dari artikel yang bersangkutan dan yang terakhir adalah *variable post* yang berisi artikel secara lengkap. Keempat variable tersebut digunakan oleh dua spider yang memuat dua metode yang berbeda, yaitu CSS *Selector* dan Xpath *Selector*.

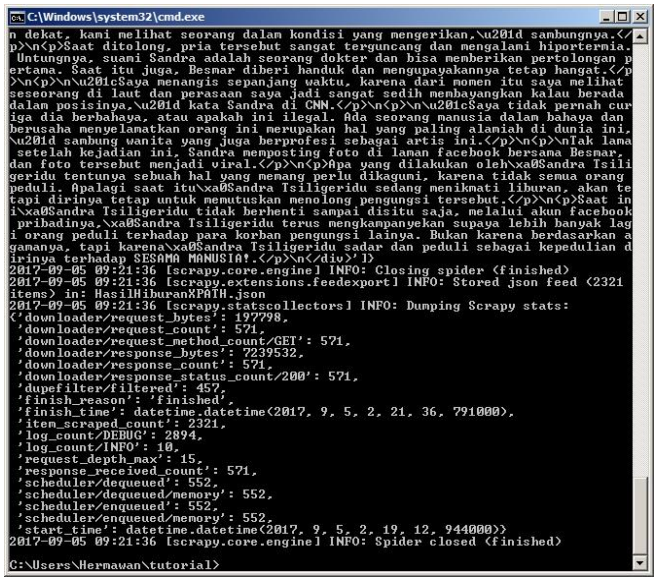
C. Pengujian

a) Tampilan program

Hasil luaran dari proses *web scraping* dalam penelitian ini disimpan kedalam file dengan ekstensi .JSON, yang kemudian disebut sebagai korpus berita komunitas, korpus berita hiburan dan korpus berita kuliner. Masing-masing korpus berita tersebut berisi hasil *web scraping* blogdetik menggunakan dua metode yaitu XPATH *Selecto*r dan CSS *Selector* yang disimpan secara terpisah berdasarkan kategori artikelnya. Hasil tampilan dari proses tersebut seperti pada gambar 2 untuk penggunaan CSS *Selector* dan gambar 3 untuk penggunaan Xpath.



Gambar 2. CSS *Selector*



Gambar 3. Xpath

b) Perbandingan Korpus Hiburan

Pada pengujian sistem terdapat empat fokus pengujian yaitu jumlah item, ukuran file .json, penggunaan memori dan waktu yang dibutuhkan untuk proses *crawling*. Hasil perbandingan proses *crawling* dengan mengunakan CSS dan Xpath untuk korpus berita hiburan ditunjukan pada Tabel 1.

Tabel 1. Perbandingan Korpus Hiburan

Korpus	Hiburan	
	CSS	XPATH
Jumlah Item	1019	2321
Ukuran (KB)	1364	7991
Penggunaan Memori	552	552
Waktu	0:00:21	0:02:24

c) Perbandingan Korpus Komunitas

Pada Tabel 2 menunjukan perbandigan fokus pengujian untuk korpus berita komunitas. Pada proses *crawling* korpus berita komunitas jumlah *item* yang didapat dengan xpath lebih banyak daripada penggunaan CSS dan berdampak pada besarnya *file* .json akan tetapi waktu yang dibutuhkan xpath untuk proses *crawling* lebih cepat dibandingkan dengan *crawling* menggunakan CSS. Pada saat dilihat hasilnya, beberapa *variable post* yang digunakan untuk menyimpan artikel berita menggunakan metode XPATH ternyata kosong. Hal ini kemungkinan disebabkan karena adanya penulisan *node* yang berbeda pada beberapa halaman web blogdetik, sehingga sistem melakukan *crawling* ke dalam *link* tersebut namun tidak bisa mengambil elemen dibawah *node* yang berbeda.

Tabel 2. Perbandingan Korpus Komunitas

Korpus	Hiburan	
	CSS	XPATH
Jumlah Item	925	1581
Ukuran (KB)	1447	5549
Penggunaan Memori	553	553
Waktu	0:03:20	0:00:18

d) Perbandingan Korpus Kuliner

Pada Tabel 3 menunjukan perbandigan untuk korpus berita kuliner. Pada proses *crawling* untuk korpus berita kuliner, jumlah *item* dan ukuran *file* yang didapat menggunakan metode XPATH lebih besar dari pada menggunakan metode CSS. Penyebabnya juga sama dengan kasus sebelumnya, yaitu terdapatnya variabel kososng yang ikut tersimpan kedalam *file*.

Dari ketiga korpus berita yang sudah didapatkan, jumlah item dan ukuran file yang didapatkan menggunakan metode XPATH lebih besar dibandingkan menggunakan metode CSS. Hal ini disebabkan karena pada saat menggunakan metode XPATH semua *node* yang berada dibawah *Selector* akan dijelajahi (*crawl*) terlepas ada atau

tidaknya variabel yang ingin disimpan, sehingga ada beberapa *item* yang tersimpan ke dalam *file* namun kosong. Selain itu, saat kita menggunakan metode XPATH, semua elemen yang berada dibawah *selector* akan ikut tersimpan. Hal ini berarti kita tidak hanya menyimpan artikel berita saja, namun juga semua kode HTML yang berada dibawah *selector*. Sehingga dibutuhkan proses lain untuk membersihkan artikel yang didapatkan dari kode HTML.

Pada proses *web scraping* menggunakan metode CSS, jumlah *item* dan *file* yang didapatkan relatif lebih kecil. Hal ini disebabkan karena hampir semua *node* yang ada pada halaman blogdetik menggunakan *style* CSS yang sama, sehingga hanya elemen dibawah *selector* yang dapat tersimpan. Selain itu, artikel yang dihasilkan juga relatif lebih bersih dari kode HTML. Hal ini disebabkan karena sistem bisa menyeleksi elemen yang dibutuhkan dengan lebih spesifik.

Tabel 3. Perbandingan Korpus Kuliner

Korpus	**Hiburan**	
	CSS	**XPATH**
Jumlah Item	316	460
Ukuran (KB)	435	1905
Penggunaan Memori	202	202
Waktu	0:02:01	0:00:08

IV. KESIMPULAN

Penggunaan metode XPATH *Selector* untuk *web scraping* situs berita menghasilkan artikel yang lebih lengkap dibandingkan dengan menggunakan metode CSS *Selector*. Hal ini ditunjukkan dengan jumlah *item* dan ukuran *file* yang didapatkan lebih besar dibandingkan metode CSS *Selector*. Namun hal ini juga menyisakan pekerjaan yang lebih banyak, karena butuh proses lain untuk menghilangkan kode HTML yang tidak diinginkan dari artikel yang dihasilkan menggunakan metode XPATH *Selector*.

Untuk penggunaan memori baik metode XPATH Selector dan CSS *Selector* tidak memiliki perbedaan yang signifikan bahkan cenderung sama. Hal ini disebabkan karena *engine* scrapy yang baik dalam penggunaan *resource*-nya, sehingga kinerjanya tidak membebani mesin, baik komputer lokal maupun server blogdetik.

Untuk waktu yang dibutuhkan pada proses *crawling* dan *scraping* secara umum metode XPATH *Selector* memiliki waktu proses yang lebih cepat daripada menggunakan metode CSS *Selector*. Pada metode XPATH, *selector* cukup mengikuti *node* pada halaman web, bukan mencari *style* halaman seperti pada metode CSS, sehingga waktu yang dibutuhkan relatif lebih singkat.

REFERENSI

[1] Akbar, S.A., Sediyonob, E. dan Nurhayati, O.D. (2015). Analisis Sentimen Berbasis Ontologi di Level Kalimat untuk Mengukur Persepsi Produk. *Jurnal Informasi Bisnis.*

[2] Kouzis-Loukas, D. (2016). *Learning Scrapy*. Birmingham-Mumbai: Packt Publishing.

[3] Hatzi, V. (2014). Web Page Download Scheduling Policies for Green Web Crawling. *22nd International Conference on Software, Telecommunications and Computer Networks (SoftCOM).*

[4] Wang, J. dan Guo, Y. (2012). Scrapy-based Crawling and User-behavior Characteristics Analysis on Taobao. *2012 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discover.*

[5] Wijaya, A.P. dan Santoso, H.A. (2016). Naive Bayes Classification pada Klasifikasi Dokumen Untuk Identifikasi Konten E-Government. *Journal of Applied Intelligent System,* pp. 48-55.