BAB 2
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Web *scraping* merupakan sebuah sebutan untuk aplikasi web yang mengambil beberapa informasi dari web lain dengan menggunakan teknik *scraping*, penerapan web *scraping* dapat digunakan di berbagai bidang. Di dalam perkembangannya telah ada beberapa orang yang melakkukan penelitian web *scraping* diantaranya sebagai berikut :

Utomo(2014) melakukan sebuah penelitian untuk mendapatkan konten dari wikipedia menggunakan teknik *reguler expresion* untuk memilah data dan data yang telah diambil kemudian akan di masukan dalam konten wordpress.

Penelitian selanjutnya dilakukan oleh Vivensius, dkk(2017) tujuan dari penelitian ini untuk melakukan penambahan data korpus dua bahasa secara otomatis ke dalam *database* korpus dua bahasa yang sudah ada, pada penelitian ini Vivensius menggunakan metode *scraping* dengan menggunakan html dom.

Penelitian ketiga dilakukan oleh Triawan(2016) penelitian ini bertujuan untuk membangun sebuah web perbadingan harga *smartphone* dari beberapa toko online. Di dalam penelitian ini mengambil data toko online berupa *merk*, harga, dan url web. Teknik pengmbilan data dengan menggunakan teknik scraping. Hasil dari penelitian ini berupa daftar informasi harga smartphone.

Penelitian keempat dilakukan oleh Kurniawati(2016), penelitian ini bertujuan untuk perangkingan dokumen al-qur'an dengan metode TF.IDF.ICSF, dengan metode ini dilakukan pencarian ayat al-alqur'an dengan mempertimbngkan jumlah kemunculan kata dalam dokumen dan kepentingan kata dalam kelas.

Kemudian penelitian terakhir dilakukan oleh Rizaldi, dan Putranto(2017) penelitian ini bertujuan menjelajahi website untuk mencari referensi-referensi korpus berita yang mana bila menemukan korpus tersebut kemudian akan dipilah data yang akan di ambil menggunakan xpath. Data hasil pemilahan akan disimpan ke dalam database. Adapun perbedaan penelitian – penelitian sebelumnya dengan penelitian yang dilakukan sekarang bisa dilihat pada tabel 2.1

Tabel 2.1 Tabel Perbandingan Penelitian

Penulis	Judul Penelitian	Objek	Teknologi	Hasil Penelitian
Mitra, Vivensius. (2017)	Rancang Bangun Aplikasi Web Scraping Untuk Korpus Paralel Indonesia – Inggris	Korpus 2 bahasa	Web scraping dengan HTML DOM	Sistem yang dapat menambah korpus 2 bahasa untuk kamus bahasa indonesia- inggris
Triawan, Deny. (2016)	Implementasi Web Crawling Perbandingan Harga Smartphone Pada Situs Jual Beli Online	Bukalapak, blibi, dan lazada	Web Crawling	Menghasilkan perbandingan harga smartphone dari 1 seri merk di beberapa toko online
Utomo, Siswo Mardi. (2014)	Web Scraping pada Situs Wikipedia	wikipedia	Web scraping dengan reguler expresi	Sistem scraping konten dengan regex
Rizaldi, Taufiq. (2018)	Pemanfaatan News Crawling Untuk Pembangunan Corpus Berita Menggunakan Scrapy dan Xpath	Blog detik	Web craling dan xpath	Sebuah sistem yang menampikan berbagai berita terbaru

Tabel 2.1 Tabel lanjutan

Penulis	Judul Penelitian	Objek	Teknologi	Hasil Penelitian
Kurniawati (2016)	term weighting berbasis indeks kelas menggunakan metode tf.idf.ics o f untuk perangkingan dokumen al-qur'an	Ayat Al-quran	Sistem temu kembali informasi	Sistem pencarian al-quran dengan memanfatkan jumlah kemunculan dalam dokumen dan tingkat kepentingan pada kelas
Usulan Penulis	Pemanfaatan web scraping dalam pengumpulan informasi lowongan kerja	Jobs.id dan topkarir	Web scraping dan tf-idf	Sistem yang dapat melakukan pencarian beberapa lowongan dalam satu proses cari.

2.1.1 Web scraping

Web scraping merupakan 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(Turland, 2010). Secara teknis *web scraping* tidak termasuk dalam bidang data mining karena data mining mengandung upaya untuk memahami pola semantik atau tren dalam kumpulan data yang besar yang telah diperoleh. Aplikassi *web scraping* berfokus hanya pada mendapatkan data dengan cara pengambalian dan ekstrasi.

Langkah-langkah melakukan web *scraping* yaitu sebagai berikut

a. Mempelajari atau observasi terhadap struktur html halaman web target.
b. Ekstraksi potongan-potongan data yang *relevan* dari halamannya.
c. Penyaringan, pemrosesan data untuk disimpan dalam *database.*

2.1.2 Php Hypertext Preprocessor

PHP merupakan singkatan dari PHP: *Hypertext Preprocessor,* secara umum PHP dikenal sebagai bahasa pemrograman script yang membuat dokumen HTML secara *on the fly* yang di eksekusi di server web(Betha Sidik, 2017). Bahasa PHP pertama kali dibuat oleh Rasmus Lerdoff, pada awalnya PHP merupakan program CGI (*common Gateway Interface*) yang dikhususkan untuk menerima input dari *form* yang di tampilkan pada *web browser*. *Software* ini kemudian disebarkan dan di lisensikan sebagai perangkat lunak *open source.*

2.1.3 Client url request library

CURL merupakan *client library* yang mendukung transfer file melalui berbagai protocol seperti FTP, HTTP, HTTPS (George, 2004). Proyek cURL merupakan proyek *open source* yang dikembangkan oleh programmer dari Swedia Daniel Stenberg dan timnya. Pustaka cURL tersedia hampir untuk semua bahasa pemrograman. Ketika cURL digunakan dengan PHP, maka dikenal dengan PHP/CURL. Nama cURL dapat diartikan gabungan dua kata client dan URL atau sebuah singkatan dari *client URL Request Library*.

2.1.4 Simple html dom

PHP simple HTML DOM merupakan sebuah wadah untuk pengembang php dan DOM. Tujuannya adalah untuk memudahkan pengembang dalam pencarian elemen DOM, dan *parsing* data(ekstrasi informasi dari tag) dengan 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.

2.1.5 Term Frequency dan Inverse Document Frequency

Term Frequency(TF) dan *Inverse Document Frequency* (TF-IDF) merupakan metode untuk menghitung bobot kata yang umum digunakan dalam sistem temu kembali informasi. Metode ini akan menghitung nilai *Term Frequency*(kemunculan kata) dan *Inverse Document Frequency* (IDF) pada setiap token atau kata di dalam dokumen(arfian. 2016). Dalam menghitung bobot kata menggunakan rumus sebagai berikut :

Rumus untuk menghitung nilai idf dapat dilihat pada persamaan (2.1)

$$IDFj = \log\left(\frac{D}{dfj}\right) \quad \text{.....(2.1)}$$

keterangan

D = jumlah semua dokumen.

Dfj = jumlah dokumen yang mengadung kata atau tf

Rumus untuk menghitung bobot kata pada dokumen dapat dilihat pada persamaan 2.2

$$Wdt = TFdt * IDFt \quad \text{.....(2.2)}$$

keterangan

d = dokumen ke-d

t = kata ke-t dari kata kunci

W = bobot dokumen ke-d terhadap kata ke-t

tf = banyaknya kata yang dicari pada sebuah dokumen

IDF = kemunculan kata di banyak dokumen

Contoh menghitung bobot tf-idf

Di dalam sebuah database terdapat beberapa judul lowongan, yaitu lowongan 1 guru bahasa inggris yogyakarta, lowongan 2 Admin Hrd, lowongan 3 admin finance & accounting, dan lowongan 4 manager - gamezone malang. Dilakukan pencariam dengan kata kunci = admin hrd yogyakarta.

Pada kata kunci ke-1 kata admin muncul pada lowongan ke-2 dan ke-3 maka nilai banyak lowongan yang mengandung kata(df) admin ada sebanyak 2 lowongan. Kata kunci ke-2 kata hrd hanya muncul pada lowongan ke-2 maka nilai df adalah 1. Kemudian kata kunci ke-3 kata yogyakarta hanya muncul pada lowongann 1, nilai df adalah 1. Setelah mendapatkan nilai df selanjutnya mencari nilai idf dan bobot lowongan terhadap kata kunci ke-i.

idfj = log(N/dfj)

idf1 = log(N/df1)

= log(4/2)

= 0301.

Kata kunci ke-1(T1) hanya muncul pada lowongan 2 dan 3 maka perhitungannya sebagai berikut

Wdt = TFdt * IDFt

W21 = TF21 * IDF1

= 1 * 0,301

= 0,301

W31 = TF31 * IDF 1

= 1 * 0,301

= 0,301

idfj = log(N/dfj)

idf2 = log(4/1)

= 0,602.

Kata kunci ke-2(T2) hanya muncul pada lowongan 2 maka perhitungannya sebagai berikut

W22 = TF22 * IDF 2

= 1 * 0,602

= 0,602

idf3 = log(4/1)

= 0,602

Kata kunci ke-1(T1) hanya muncul pada lowongan 2 dan 3 maka perhitungannya sebagai berikut

W21 = TF21 * IDF 1

= 1 * 0,301

= 0,301

dengan demikian dapat diperoleh nilai bobot(w) untuk setiap kata kunci pada masing-masing lowongan.

Term(kata)	Kemunculan kata				df	d/df	IDF	Bobot kata pada dokumen			
	L1	L2	L3	L4				L1	L2	L3	L4
admin	0	1	1	0	2	4/2	0,301	0	0,301	0,301	0
hrd	0	1	0	0	1	4/1	0,602	0	0,602	0	0
yogyakarta	1	0	0	0	1	4/1	0,602	0,602	0	0	0
Jumlah bobot tiap lowongan								0,602	0,903	0,301	0