JATIKOM: Jurnal Aplikasi dan Teori Ilmu Komputer | ISSN: 2615-577X (Online)
https://ejournal.upi.edu/index.php/JATIKOM

Implementasi Penerapan Metode *Scraping* pada Pembuatan *Curiculum Vitae*

Irfan Haydar Rachman[1], Rani Megasari[2], Eddy Prasetyo Nugroho[3]

Program Studi Ilmu Komputer Departemen Pendidikan Ilmu Komputer Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam Universitas Pendidikan Imdonesia
Bandung Indonesia

[1]irfan.haydar.rachman@studen.upi.edu, [2]megasari@upi.edu, [3]eddypn@upi.edu

Abstrak— ***Curriculum vitae*** **merupakan dokumen yang memberikan gambaran rinci tentang pengalaman, kualifikasi dan prestasi seseorang, terutama hal-hal yang berhubungan dengan akademis. CV berisi tentang Identitas pribadi, riwayat pendidikan, bidang/ spesifikasi keilmuan yang ditekuni, matakuliah yang diampu dalam 3 tahun terakhir, kegiatan pengabdian masyarakat yang dilakukan 3 tahun terakhir, buku teks yang diterbitkan oleh penerbit dalam 3 tahun terakhir, seminar dalam bidang keilmuan, kerjasama yang pernah dilakukan, Setiap dosen diharuskan untuk memiliki *Curiculum Vitae*. seorang dosen yang aktif dalam jangka 1 tahun dapat melakukan banyak kegiatan khususnya dalam bidang akademis, oleh karena itu setiap dosen harus memiliki CV yang terbaru untuk kebutuhan kegiatan dan lain-lain. Antisipasi yang dilakukan adalah membuat sebuah system yang akan membuat CV dengan cara Web Scraping dari sumber yang valid, aplikasi ini bekerja sesuai dengan kebutuhan dan menghasilkan CV dengan data yang dibutuhkan. Teknik Scraping yang dipakai yaitu HTML DOM Parser**

Kata Kunci: *CV, Web Scraping, DOM Parser, HTML.*

Abstract— **Abstract in English…**

Keywords: *CV, Web Scraping, DOM Parser, HTML.*

I. PENDAHULUAN

Curriculum vitae atau yang lebih dikenal dengan istilah CV merupakan istilah Latin yang berarti "cerita kehidupan". CV adalah dokumen yang memberikan gambaran rinci tentang pengalaman, kualifikasi dan prestasi seseorang, terutama hal-hal yang berhubungan dengan akademis. Dokumen ini juga bisa berisi informasi pribadi seperti status pernikahan, kewarganegaraan, tanggal lahir, dan bahkan foto. CV ditulis sebagai template tanpa perubahan, kecuali jika ada kualifikasi atau pencapaian baru yang ingin ditambahkan ke dalam daftar.

Sangat penting bahwa CV itu akurat, karenanya pentingnya menuliskan prestasi, penghargaan, dan pekerjaan yang telah dilakukan. Jika CV yang ditulis kurang tepat, kemungkinan akan merusak kredibilitas, karena pemeriksaan referensi dan pencarian mudah dilakukan untuk memverifikasi informasi, dan bahkan kesalahan yang tidak disengaja akan dilihat dengan jelas. [1].

Selain memperbarui CV, ketika telah mencapai tujuan tertentu atau selama interval setengah tahun atau tahunan, CV harus dipoles agar relevan dengan pekerjaan tertentu atau promosi yang dicari. Pastikan untuk menggunakan periksa ejaan dan mengoreksi setiap halaman dengan cermat setiap kali mengirimkannya. Singkatnya, penting untuk mempertahankan CV yang terbaru dan akurat, yang dapat dengan mudah disesuaikan karena keadaan dan CV perlu diubah.

Web scraping (web harvesting atau web data extraction) adalah data scraping yang digunakan untuk mengekstrak data dari situs web. Web scraping software dapat mengakses *World Wide Web* (WWW) secara langsung dengan menggunakan Hypertext Transfer Protocol (HTTP), atau melalui web browser. Ini adalah bentuk penyalinan, di mana data spesifik dikumpulkan dan disalin dari suatu web, dan disimpan ke database lokal atau spreadsheet pusat, yang nanti nya dapat kita gunakan lagi di proses selanjutnya.

Pada penelitian yang dilakukan oleh Vivensius Mitra, Herry Sujaini, Arif Bijaksana Putra Negara, Sistem mampu menghasilkan dokumen korpus paralel melalui proses scraping dengan metode HTML DOM dari website Berita dua Bahasa dengan alamat URL (http://www.berita2bahasa.com/) dan mampu menghasilkan dokumen yang berisi kumpulan berita dan artikel Bahasa Indonesia sebagai sumber dan Bahasa Inggris sebagai terjemahan. Pengumpulan korpus paralel dengan Aplikasi Web Scraping dengan Metode HTML DOM sangat berpengaruh pada spesifikasi perangkat keras dan kecepatan internet yang digunakan meskipun waktu yang dibutuhkan cukup lama namun sesuai pada jumlah korpus paralel yang diperoleh, dan sistem ini jauh lebih praktis dibandingkan dengan proses manual dalam pengumpulan korpus parallel [2].

Sedangkan pada penelitian lain, Proses untuk memisahkan konten utama halaman situs dengan bagian-bagian yang tidak berhubungan dengan isi disebut dengan scraping. Dengan teknik ini konten utama dari suatu halaman situs dapat diekstrak, dikoleksi dan selanjutnya

dapat diproses oleh proses pengindekan. Sistem ini adalah perangkat lunak berbasis web dengan tujuan melakukan pengambilan isi dari konten halaman web. Hal-hal yang dapat diwujudkan dalam sistem ini diantaranya Sistem dapat secara otomatis mengekstrak konten utama dari suatu halaman web, Dalam penelitian ini digunakan halaman dokumen pada situs resmi sebuah produk makanan dengan merk Bango, Pengambilan data/crawling Uniform Resource Locator (URL) pada situs resmi sebuah produk makanan merk Bango menggunakan aplikasi sphider, Hasil scraping resep disimpan dalam basisdata, Sistem ini dapat memproduksi data resep dengan format XML (eXtensible Markup Language). Aplikasi diintegrasikan dalam bentuk plugin CMS wordpress yang dapat diunduh di secara bebas. Sistem diimplementasikan secara online menggunakan sebuah situs yang telah disiapkan. Teknik web scraping dapat digunakan untuk mengambil konten resep pada situs pada berbagai situs yang memuat resep masakan. Penyimpanan resep ke dalam basisdata, mempermudah transformasi data ke bentuk lainnya [3].

Oleh karena itu penulis akan mengimplementasikan metode web scraping untuk pembuatan *Curiculum Vitae* (CV), agar mempunyai CV yang relevan dengan prestasi terbaru yang telah dicapai, penulis akan membuat beberapa opsi template, agar CV yang dihasilkan tidak terlalu monoton dan memiliki hasil yang berbeda, data yang akan diambil dan digunakan untuk bahan pembuatan CV ini berasal dari website yang telah ditentukan.

II. PEMBAHASA PENELITIAN

A. *Curriculum Vitae*

Curriculum vitae atau CV adalah dokumen yang memberikan gambaran rinci tentang pengalaman, kualifikasi dan prestasi seseorang, terutama hal-hal yang berhubungan dengan akademis. Dokumen ini juga bisa berisi informasi pribadi seperti status pernikahan, kewarganegaraan, tanggal lahir, dan bahkan foto. CV ditulis sebagai template tanpa perubahan, kecuali jika ada kualifikasi atau pencapaian baru yang ingin ditambahkan ke dalam daftar. CV biasanya berisi minimal 2 halaman tapi bisa jauh lebih banyak saat dimaksudkan untuk menjadi dokumen terperinci.

Sangat penting bahwa CV itu akurat, karenanya pentingnya menuliskan prestasi, penghargaan, dan pekerjaan yang telah dilakukan. Jika CV yang ditulis kurang tepat, kemungkinan akan merusak kredibilitas, karena pemeriksaan referensi dan pencarian mudah dilakukan untuk memverifikasi informasi, dan bahkan kesalahan yang tidak disengaja akan dilihat dengan jelas [1]. Dalam lingkungan yang sangat kompetitif, CV harus disesuaikan dan diorganisir untuk menyoroti pencapaian individu (Kelsky, 2012). Janji profesional, publikasi, penghargaan dan penghargaan, hibah, dan konferensi yang diundang Setiap presentasi semua melibatkan peer review dan kompetisi. Ini harus disebutkan sejak awal di CV. Meskipun kegiatan *peer-review* ini mungkin dianggap baik untuk promosi di komunitas akademik, pendidikan dan layanan masyarakat memberikan kontribusi yang berharga bagi budaya dan pengembangan profesional dan tempat kerja [4]. Dengan demikian, itu tergantung pada konteks bagaimana menyeimbangkan pencapaian pribadi yang baik dalam kaitannya dengan kontribusi kepada profesi atau organisasi masyarakat. Berkonsultasi dengan dan mencari umpan balik dari mentor atau rekan tepercaya dapat membantu mengklarifikasi masalah yang berkaitan dengan konten dan penekanan. Pada akhirnya, CV memberikan kepada pembaca potret kemampuan, prestasi, kekuatan, dan potensi; oleh karena itu, dapat memakan banyak waktu untuk tugas ini (Leung & Robson, 1990).

B. *Web Scraping*

Web scraping adalah sebuah proses yang memanfaatkan dokumen berbentuk semi – structured yang didapatkan dari internet, yang dimana dokumen tersebut berbentuk sebuah halaman website yang dibangun oleh bahasa markup seperti HTML ataupun XHTML yang kemudian dianalisis untuk mendapatkan informasi yang berguna yang dapat dilakukan untuk konteks lain.

C. *Teknik-Teknik Web Scraping*

1) Parsing HTML

   Merupakan salah satu teknik yang paling banyak dipakai dalam web parsing. Biasanya parsing HTML dilakukan melalui bahasa JavaScript lalu menarget halaman HTML linear serta nested. Metode ini termasuk cepat untuk mengidentifikasi script HTML di website, yang mungkin saja dilakukan secara manual. Script ini kemudian dipakai untuk mengekstraksi text, links, dan data.

2) Parsing DOM

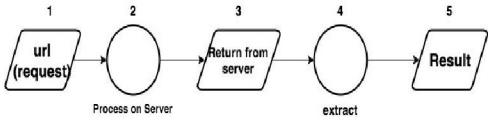
   Konten, style, serta struktur file XML didefinisikan dalam DOM, singkatan dari Document Object Model. Scrapers yang ingin mengetahui cara kerja internal pada halaman web lalu mengekstrak skrip yang berjalan di dalamnya, biasanya memilih untuk melakukan web scraping melalui teknik parsing DOM. Node spesifik dikumpulkan memakai parser DOM serta alat-alat seperti XPath membantu proses scraping pada sebuah halaman web.

3) XPath

   XML Path Language atau lebih dikenal dengan XPath, merupakan bahasa query yang bekerja pada dokumen XML. Karena dokumen XML biasa disusun menggunakan struktur pohon (tree structure), XPath dapat digunakan untuk menavigasi struktur dokumen tersebut dengan cara memilih nodes berdasarkan berbagai parameter. XPath juga dapat digunakan bersamaan dengan teknik DOM parsing dalam mengesktrasi seluruh halaman website lalu menampilkannya di website lain.

D. HTML DOM Parsel

HTML DOM (Document Object Model) adalah kumpulan obyek-obyek pada elemen HTML. HTML (HyperText Markup Language) sendiri bisa berarti sebuah bahasa untuk membuat situs web, dan menampikan informasi pada suatu situs penjelajah web internet dalam format hypertext ASCII agar dapat menghasilkan tampilan yang terintegrasi.



Gambar. 1 Tahapan Web Scraping

Menurut Gambar diatas berikut penjelasan mengenai cara kerja Web Scraping dengan metode HTML DOM parse :

1) *Request* url yang dijadikan target.
2) *Request* diproses oleh server target.
3) Hasil dari *request* url (hasilnya adalah text dengan format HTML).
4) Ekstrak data (mengambil data yang diperlukan dari tahap ke-3).
5) Hasil yang ekstrak (menentukan output yang diingkan).

Contoh Penggunaan Simple HTML DOM:

```
<?php

require_once 'simple_html_dom.php';

//menguraikan sebuah halaman situs http:// atau https://
$html = file_get_html("http://hieppies.blogspot.com");

//menguraikan pada file lokal
$html = file_get_html("index.html");

//menguraikan pada file lokal yang didalam direktori
$html = file_get_html("/menu/index.html");

//menguraikan string pada kode HTML
$html = str_get_html("<html>
                <body>
                  <p>Hi, Cantik!.</p>
                  <p>Lagi ngapain?.</p>
                </body>
                </html>");
$elements = $html->find("p");

//menampilkan string pada paragraf pertama
echo $elements[0]->plaintext;

//menampilkan string pada paragraf kedua
echo $elements[1]->plaintext;

//menambahkan atribut class pada paragraf pertama
$elements[0]->class = "nama_class";

//menampilkan HTML yang ditambah class
echo $html->save();
?>
```

HTML yang dihasilkan dari perintah $html->save();

```
<html>
<body>
   <p class="nama_class">Hi, Cantik!.</p>
   <p>Lagi ngapain?.</p>
</body>
</html>
```

Penyeleksian (Selector) lainnya.

Berikut adalah beberapa contoh lain dari penyeleksian melalui atribut. Jika anda terbiasa menggunakan jQuery, ini sangat mudah bagi anda.

```
<?php

//ambil elemen index pertama dengan atribut id="foo"
$single = $html->find("#foo", 0);

//ambil semua tag elemen yang beratribut class="foo"
$collection = $html->find('.foo');

//ambil semua tag elemen "a"
$collection = $html->find('a');

//ambil semua tag elemen a yang didalam tag elemen h1
$collection = $html->find('h1 a');

//ambil semua tag elemen img yang beratribut title="himom"
$collection = $html->find('img[title=himom]');

   ?>
```

III. HASIL PEMBAHASAN

A. Hasil

Penelitian dimulai dengan melakukan studi literatur mengenai teori-teori dan kajian yang berkaitan dengan Teknik Web Scraping dan Curriculu Vitae.

Sangat penting bahwa CV itu akurat, karenanya pentingnya menuliskan prestasi, penghargaan, dan pekerjaan yang telah dilakukan. Jika CV yang ditulis kurang tepat, kemungkinan akan merusak kredibilitas, karena pemeriksaan referensi dan pencarian mudah dilakukan untuk memverifikasi informasi, dan bahkan kesalahan yang tidak disengaja akan dilihat dengan jelas.

Pada Penelitian ini, metode yang digunakan dalam perangkat lunak adalah Web Scraping dengan Teknik Parsing DOM. Teknik parsing DOM bekerja dengan cara mencari kata kunci dari sebuah tag HTML atau kata sesuai apa yang ingin ditampilkan.

Data yang digunakan dalam penelitian ini adalah data yang bersumber dari data dosen Universitas Pendidikan Indonesia, yang bersumber pada https://dosen.upi.edu.

B. Tahapan Scraping

Tahapan Scraping Dalam penelitian ini yang pertama menentukan terlebih dahulu url yang akan dijadikan target. Di dalam penelitian ini sumber data yang akan diambil adalah dari url https://dosen.upi.edu .

Dalam pembuatan aplikasi ini digunakan library yang bernama html_dom. Langkah – Langkah nya seperti berikut:

1) Memanggil library dan mengekstrak HTML dari ulr yang dimaksud

```
//memanggil library pasring html
include('simple_html_dom.php');

// Create DOM from URL or file
$html = file_get_html('http://dosen.upi.edu//umum/biografi/summary/' $nip);
```

Gambar. 2 Memanggil *Libraly*

2) Untuk mengambil data yang dimaksud, kita harus mengetahui element – element nya dengan cara inspect element di halaman halaman browsing atau page source untuk melihat keseluruhan.

```
::before
<div class="col-md-2">…</div>
<div class="col-md-10">
  <table class="table table-striped"> == $0
    <tbody>
      <tr>
        <td>
          <b>Nama Lengkap</b>
        </td>
        <td>: Prof. Dr. MUNIR, M.IT.</td>
      </tr>
      <tr>
        <td>…</td>
        <td>: ICT for Education</td>
      </tr>
      <tr>…</tr>
      <tr>…</tr>
```

Gambar. 3 *Inspect Element* url Sumber

Dilihat dari gambar 3.2 data berada di dalam table, karena tidak ada id di dalam table, jadi kita harus mencari keyword untuk mengambil semua data yang dimaksud

3) Langkah berikutnya mencari table di dalam HTML yang sudah di ekstrak .

```
// Mencari data melalui table yg ada di web upi
foreach($html->find('table') as $element){
```

Gambar. 4 Mencari Tag HTML

Dari gambar diatas adalah perintah untuk mencari table di dalam HTML.

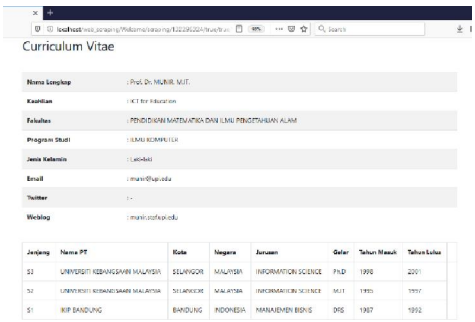
4) Karena table nya tidak memiliki id jadi kita harus memakai keyword untuk mencari data yang dimaksud.

```
//untuk menampilkan riwayat hidup
if((strpos($element, "Nama Lengkap") != null && ($riwayat_hidup == "true"))){
    //echo $element . '<br>';
    $data['pendidikan'][] = $element . '<br>';
}
```

Gambar. 5 Mencari *Keyword*

Menurut gambar 3.4 mencari data yang dimaksud menggunakan syntax strops. Lalu data disimpan di dalam array untuk nanti ditampilkan

5) Hasil dari Scraping ada di gambar 3.4 ditampilkan sesuai pilihan dari user.



Gambar. 6 Hasil *Scraping*

IV. KESIMPULAN

Teknik Web Scraping sangat membantu para dosem untuk membuat CV yang diinginkan dan sesuai kebutuhan. Data yang diambil merupakan data yang valid karena bersumber dari website UPI. Hasil scraping CV dapat di download berupa pdf.

Teknik HTML DOM Parser terbukti sangat memudahkan untuk mengekstrak data dan mencari data yang diinginkan berdasarkan keyword atau tag HTML. Untuk menggunakan Scraping ini dibutuhkan koneksi internet yang sangat lancar.

DAFTAR PUSTAKA

[1] M. &. H. J. Cleary, 2013. "Keeping Your *Curriculum vitae* Up To Date," *Journal of Psychosocial Nursing and Mental Health Services,* vol. 51, no. 6, pp. 4-5.

[2] V. S. H. A. B. &. N. P. Mitra, 2017. "untuk Korpus Paralel Indonesia - Inggris dengan Metode HTML DOM," *Jurnal Sistem Dan Teknologi Informasi (JUSTIN),* vol. 5(1), p. 1–6.

[3] S. &. U. M. S. Wibisono, 2013."Perancangan Aplikasi Web Scraping Untuk Koleksi Konten Resep Masakan Tradisional Jawa Berbasis XML," vol. 1, pp. 1-7.

[4] D. L. F. C. F. P. R. E. S. V. &. S. P. Quoc, 2015. "UniCrawl: A Practical Geographically Distributed Web Crawler. Proceedings," *IEEE 8th International Conference on Cloud Computing,* p. 389–396.