IMPLEMENTASI WEB SCRAPING UNTUK PENGAMBILAN DATA REKOMENDASI FILM PADA IMDB

WEB SCRAPING IMPLEMENTATION FOR
FILM RECOMMENDATION DATA COLLECTION ON IMDB

Ivana Elfirdaus[1*], Eka Dyar Wahyuni[1]
*E-mail: ivanaelfirdaus@gmail.com

[1]Sistem Informasi, Fakultas Ilmu Komputer, UPN "Veteran Jawa Timur"

Abstrak

Teknologi yang terus berkembang saat ini memberikan banyak dampak pada masyarakat. Penerapan teknologi yang tepat dapat mengatasi berbagai permasalahan yang ada. Permasalahan tersebut salah satunya mengenai terbatasnya akses data di beberapa platform untuk pengambilan data. Untuk mengatasi permasalahan tersebut, dapat dilakukan berbagai macam teknik integrasi, terutama pada level *user interface*. Penelitian ini menerapkan integrasi level *user interface* menggunakan teknik web *scraping* dari situs IMDb dengan menggunakan ekstensi dari Google Chrome yaitu web *scraper* untuk menghasilkan file CSV. Hasil dari web *scraping* ini berupa data detail dari rekomendasi film terpopuler yang selanjutnya dilakukan *transform* data menggunakan Microsoft Excel dan kemudian dilakukan *import* data ke dalam MySQL. Data yang didapatkan dalam penelitian ini kemudian divisualisasi dalam bentuk *website* dengan mempergunakan *template* bootstrap.
Kata kunci: *implementasi, web scraping, data, IMDb*

Abstract

Technology that continues to develop today has a lot of impact on society. The application of the right technology can overcome various existing problems. One of these problems is regarding limited data access on several platforms for data collection. To overcome these problems, various kinds of integration techniques can be carried out, especially at the user interface level. This study applies user interface level integration using web scraping techniques from the IMDb site by using an extension from Google Chrome, called the web scraper to generate CSV files. The results of this web scraping are detailed data from the most popular film recommendations which are then transformed using Microsoft Excel and then load the data into MySQL. The data obtained in this study is then visualized in to a website using the bootstrap template.
Keywords: *implementation, web scraping, data, IMDb*

1. PENDAHULUAN

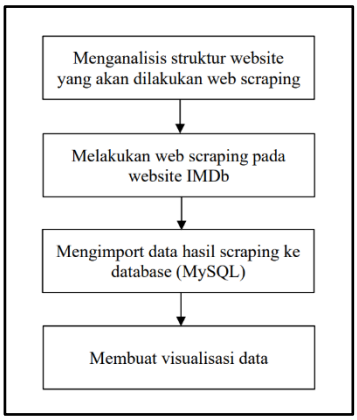
Perkembangan teknologi saat ini sangat berpengaruh dalam aktivitas keseharian masyarakat. Hal tersebut mengharuskan masyarakat untuk bisa beradaptasi dengan teknologi saat ini. Dengan adanya penerapan teknologi, seluruh aktivitas masyarakat akan sangat dimudahkan. Tetapi disisi lain, penerapan teknologi juga masih menimbulkan beberapa masalah. Permasalahan yang terjadi salah satunya adalah adanya keterbatasan mengakses data yang disebabkan oleh adanya beberapa platform yang tidak mengizinkan untuk dilakukan pengambilan data yang diperlukan [1].
Penerapan teknologi yang sedang *trend* saat ini adalah penggunaan internet. Menurut survei pada Asosiasi Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) mencatat bahwa pada tahun 2023 penggunaan internet mencapai 215.626.156 jiwa dari total populasi yang sebesar

275.773.901 jiwa. Kenaikan jumlah pengguna tersebut sebesar sebesar 1,17 persen [2]. Salah satu pemanfaatan dari internet untuk hiburan saat ini adalah pencarian informasi mengenai rekomendasi film menggunakan *Search engine* (mesin pencarian) seperti Google. *Search engine* ini mempergunakan bot yang akan menelusuri dan mengumpulkan informasi dari beberapa *website* secara periodik. Proses dari pengambilan informasi pada *website* tersebut dapat disebut dengan "web scraping" [3].

Web Scraping merupakan sebuah metode yang dilakukan dalam mengekstraksi data spesifik dari suatu web. Data yang sudah terekstraksi dapat dikumpulkan pada Google Sheets atau Microsoft Excel [4]. Sedangkan menurut teori, *web scraping* merupakan sebuah cara dalam pengumpulan data dengan menggunakan metode yang berbeda dengan penggunaan API (*Application Programming Interface*) [5]. *Web scraping* merupakan salah satu implementasi dari integrasi level *user interface*. *Web scraping* ini sering dikenal sebagai *screen scraping* atau dapat juga disebut dengan *intelligent*, *automated*, or *autonomous agents.* Teknik ini berfokus terhadap cara memperoleh data melalui pengambilan dan ekstraksi data dengan ukuran yang beragam [3].

Penerapan web *scraping* ini dapat diimplementasikan untuk mengumpulkan informasi tertentu dari sebuah *website* yang nantinya diolah dan ditampilkan dalam bentuk lain berupa informasi untuk para penggunanya [6]. Keuntungan dengan menggunakan web *scraping* ini yaitu waktu yang cukup efisien, memungkinkan pengambilan data secara teratur dengan jangka waktu yang singkat, dan dapat memperoleh jumlah informasi yang lebih banyak [7]. Menimbang keuntungan tersebut, penelitian ini berfokus pada integrasi level user interface untuk mengambil beberapa data yang terdapat pada situs *website* IMDb menggunakan teknik *web scraping*. Dalam penelitian ini, tujuan dilakukan *web scraping* adalah untuk mendapatkan data lengkap mengenai rekomendasi film populer dari situs *website* IMDb dan divisualisasikan menggunakan template bootstrap berupa *website* yang lebih mudah untuk dipahami [8]. *Web scraping* dilakukan dengan menggunakan ekstensi dari Google Chrome yaitu *web scraper.* Data-data yang sudah berhasil didapatkan, nantinya akan divisualisasikan dalam bentuk tampilan *website*. Harapan peneliti terhadap hasil dari penelitian ini dapat memberi kontribusi akademis yaitu cara menggunakan *web scraping* dalam melakukan ekstraksi data khususnya pada *website* IMDb dan memberi kontribusi berupa kebijakan dalam hal pencarian data melalui internet oleh peneliti sendiri [9].

2. METODOLOGI



Gambar 1. Alur Penelitian

Alur pada penelitian ini dapat dilihat pada Gambar 1 yang terdiri dari 4 tahapan. Tahapan yang pertama, peneliti melakukan analisis struktur pada *website* yang akan dilakukan proses *web scraping*. Hal tersebut perlu dilakukan terlebih dahulu untuk menghindari adanya data yang tidak dapat dilakukan proses web scraping. Tahapan kedua, mengimplementasikan *web scraping*. Peneliti memilih pengambilan data dari *website* IMDb (https://www.imdb.com/?ref_=nv_home)

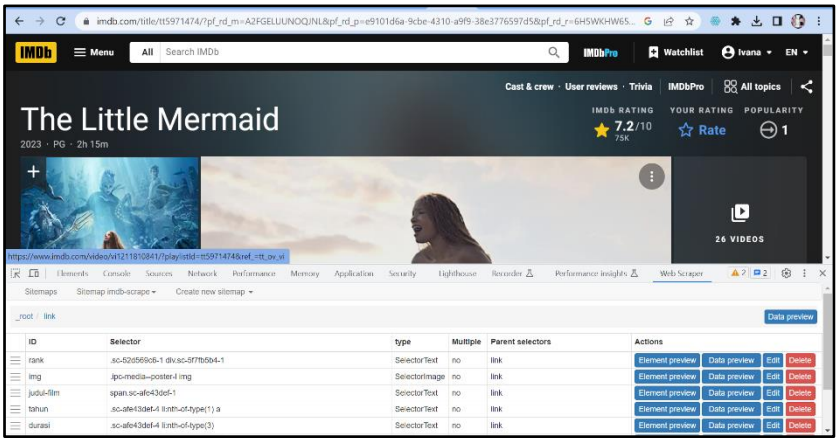
berupa judul film, tahun, durasi, kategori, dan rating dengan menggunakan ekstensi dari Google Chrome yaitu *web scraper*. Hasil dari proses web *scraping* tersebut adalah berupa data XLSX atau CSV yang diperoleh dari halaman *website* dalam bahasa *markup* yaitu HTML atau XHTML [10]. Hasil *scraping* tersebut selanjutnya dilakukan proses transformasi data menggunakan Microsoft Excel dengan menghapus beberapa kolom yang tidak diperlukan pada visualisasi nantinya. Proses ini merupakan salah satu jenis dari penerapan integrasi level data yang bertujuan untuk melakukan penyesuaian sumber data dengan struktur dan format yang dibutuhkan oleh sistem [11].

Pengambilan data yang sudah berhasil dilakukan sebelumnya akan dilanjutkan pada tahapan yang ketiga yaitu, *import* hasil *web scraping* ke dalam database MySQL. Setelah hasil *scraping* sudah berhasil masuk pada database MySQL, dilanjutkan dengan tahapan terakhir pada penelitian ini yaitu visualisasi data dalam bentuk halaman *website*. Visualisasi pada *website* tersebut dibuat menggunakan bahasa pemrograman PHP yang terhubung dengan *database* MySQL. Mengkonfigurasi koneksi ke database yang tersimpan pada file indexwatch.php dengan menggunakan tampilan tabel dari *template bootstrap.* Pembuatan tampilan web dengan menggunakan *Bootstrap* dapat mempermudah dalam menyesuaikan lebar dari medianya yang membuat tampilan web akan rapi ketika dibuka dengan berbagai media lainnya seperti handphone, tablet, laptop ataupun PC desktop. Adanya hal tersebut dapat membuat tampilan tidak akan terganggu pada resolusi dari layar yang digunakan [12].

3. HASIL DAN PEMBAHASAN

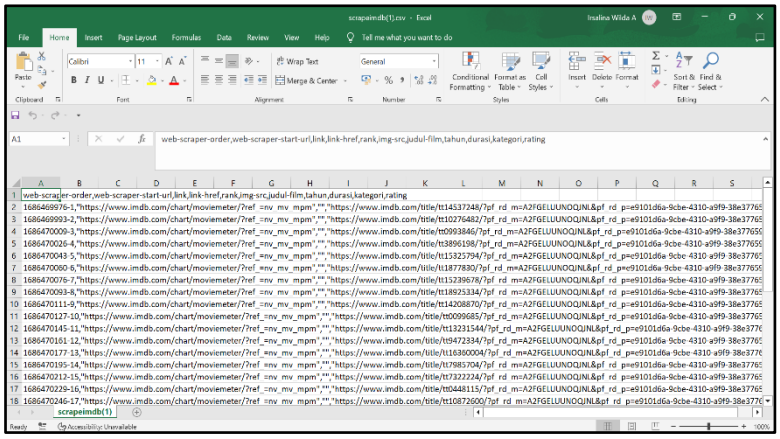
3.1 Hasil Proses Web *Scraping*

Data yang diperlukan terdapat pada halaman di situs *website* resmi IMDb (https://www.imdb.com/?ref_=nv_home). Selanjutnya penyesuaian data apa saja yang akan diambil dan selector apa saja yang diperlukan. Kemudian dilanjutkan dengan *scraping* data seperti pada Gambar 2.



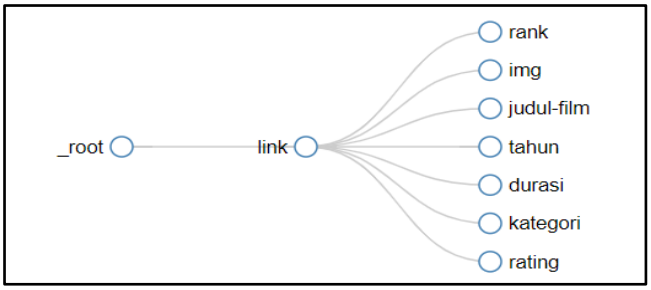
Gambar 2. Proses Web *Scraping*

Hasil *scraping* tersebut memberikan data yang ditunjukkan pada Gambar 3, bahwa data tersebut didapatkan dari otomatisasi pengambilan data pada setiap judul film yang dilengkapi dengan detail lainnya. Data yang berhasil diambil melalui proses *scraping* kemudian dilakukan ekspor data berupa file CSV.



Gambar 3. Data CSV Film Rekomendasi IMDb

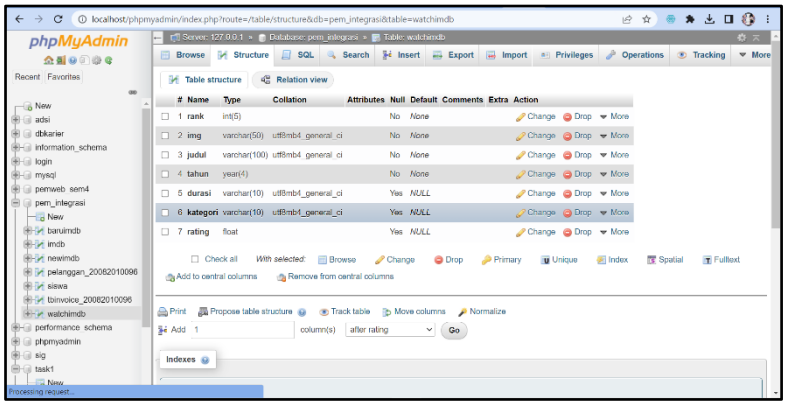
Data CSV tersebut didapatkan dari hasil dan skema yang digambarkan pada *Selector Graph*. *Selector Graph* berbentuk seperti pohon yang memiliki cabang pada tiap sub-nya yang ditunjukkan pada Gambar 4.



Gambar 4. *Selector Graph*

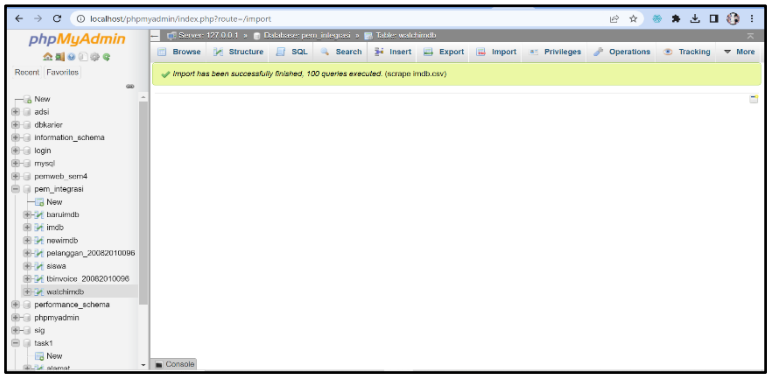
3.2 Hasil Proses *Import* Data

Sebelum dilakukan visualisasi pada *website* yaitu membuat *database* terlebih dahulu di MySQL dengan struktur tabel seperti pada gambar 5. Struktur tersebut disesuaikan dengan isi data yang diperlukan dalam visualisasi data.



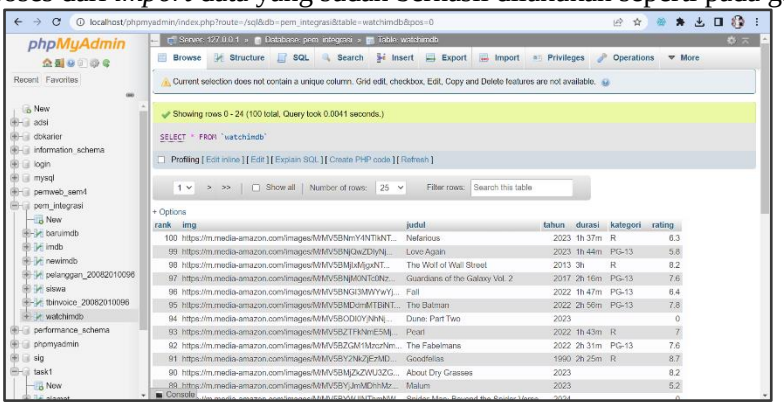
Gambar 5. Struktur Tabel MySQL

Setelah database dibuat, selanjutnya dilakukan *import* data CSV yang sudah didapatkan dari hasil *scraping* sebelumnya ke dalam tabel yang sudah dibuat.



Gambar 6. Proses *Import* Data CSV ke MySQL

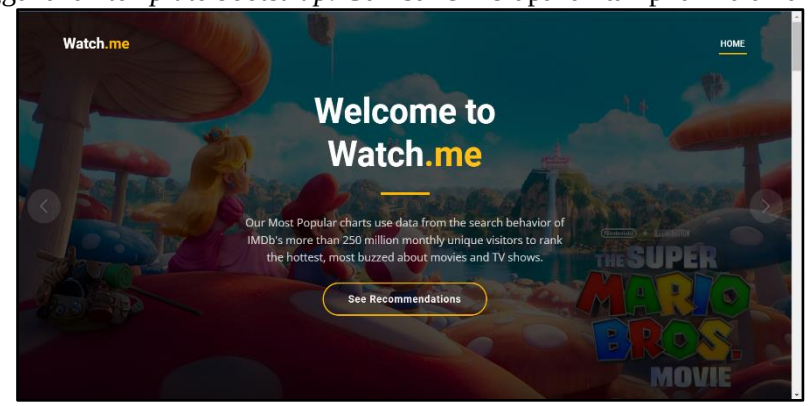
Berikut hasil proses dari *import* data yang sudah berhasil dilakukan seperti pada gambar 7.



Gambar 7. Hasil Import Data ke MySQL

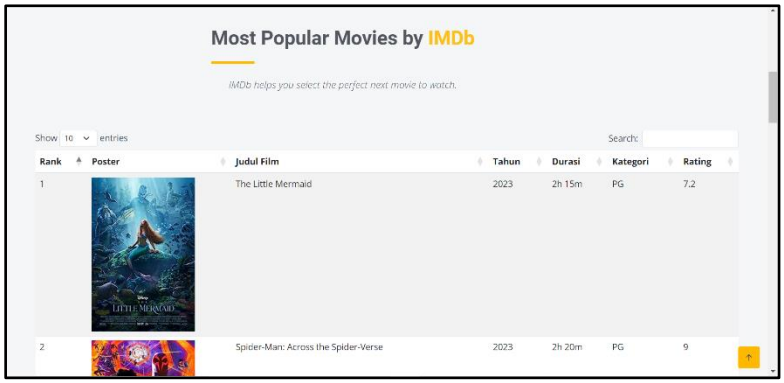
3.3 Hasil Visualisasi

Pembuatan tampilan halaman *website* dibuat terlebih dahulu sebelum melakukan visualisasi data dengan menggunakan *template bootstrap.* Gambar 8 merupakan tampilan halaman.



Gambar 8. Tampilan *Website*

Setelah pembuatan *website* berhasil dilakukan, selanjutnya membuat visualisasi data yang berasal dari *database* ke dalam halaman *website*. Data tersebut selanjutnya disesuaikan pada tabel berdasarkan rank, poster, judul film, tahun, durasi, kategori, dan rating seperti gambar 9.



Gambar 9. Tampilan Halaman Visualisasi Data

4. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengimplementasikan *web scraping* terhadap situs IMDb, kemudian menampilkan hasil *scrapping* tersebut dalam suatu *website*. Teknik *scrapping* ini memberikan kemudahan dalam mengambil data yang dibutuhkan pada *website* secara otomatis dalam ukuran kecil hingga jumlah besar, jika *website* yang dituju tidak memberikan mekanisme API untuk mengakses datanya. Dalam memanfaatkan *web scraping* ini juga perlu diperhatikan etika dan hukum TI. Beberapa situs web memiliki kebijakan yang melarang *scraping*. Sebelum dilakukan *scrapping*, penting untuk memahami aturan dan kebijakan yang berlaku untuk setiap situs web yang ingin dilakukan *scraping.* Selain itu, diperlukan juga tanggungjawab dan taat hukum agar pemanfaatan *scrapping* ini membawa dampak yang positif.

5. DAFTAR RUJUKAN

[1] A. R. Salsabila, Muhammad Daffa, M. K. Happy Maulana, And E. D. Wahyuni, "Implementasi Teknik Web Scraping Untuk Menampilkan Data Tim English Premier League," *Prosiding Seminar Nasional Teknologi Dan Sistem Informasi (Sitasi)* , Vol. 2, Sep. 2022.

[2] "Survei Apjii Pengguna Internet Di Indonesia Tembus 215 Juta Orang," *Asosiasi Penyelenggara Jasa Internet Indonesia (Apjii)*, Mar. 10, 2023. Https://Apjii.Or.Id/Berita/D/Survei-Apjii-Pengguna-Internet-Di-Indonesia-Tembus-215-Juta-Orang (Accessed Jun. 09, 2023).

[3] Ahmad Josi, L. A. Abdillah, And Suryayusra, "Penerapan Teknik Web Scraping Pada Mesin Pencari Artikel Ilmiah," *Jurnal Sistem Informasi (Sisfo)*, Vol. 5, Oct. 2014.

[4] "Apa Itu Web Scraping? Cara Kerja, Teknik & Kegunaannya," *Jogjahost*, Feb. 23, 2023. Https://Www.Jogjahost.Co.Id/Blog/Apa-Itu-Web-Scraping/ (Accessed Jun. 15, 2023).

[5] Y. Sahria, "Implementasi Teknik Web Scraping Pada Jurnal Sinta Untuk Analisis Topik Penelitian Kesehatan Indonesia," *Proceeding Of The 11th University Research Colloquium 2020: Bidang Sains Dan Teknologi* , May 2020.

[6] A. Sopian, A. Dharmalau, And Alpindo, "Pemanfaatan Teknik Web Scraping Python Untuk Sistem Pencarian Produk Di Toko Online," *Jurnal Elektro & Informatika Swadharma (Jeis)*, Vol. 2, Jul. 2022.

[7] A. A. Maulana, A. Susanto, And D. P. K., "Rancang Bangun Web Scraping Pada Marketplace Di Indonesia," *Journal Of Information System*, Vol. 4, May 2019.

[8] A. S. Yondra, D. Triyanto, And S. Bahri, "Implementasi Web Scraping Untuk Mengumpulkan Informasi Produk Dari Situs E-Commerce Dan Marketplace Dengan Teknik Pemrosesan Paralel," Vol. 10, 2022.
[9] Mohammad Djufri, "Penerapan Teknik Web Scraping Untuk Penggalian Potensi Pajak (Studi Kasus Pada Online Market Place Tokopedia, Shopee Dan Bukalapak)," *Jurnal Bppk : Badan Pendidikan Dan Pelatihan Keuangan*, Vol. 13, Dec. 2020.
[10] D. F. Setiawan, T. Tristiyanto, And A. Hijriani, "Aplikasi Web Scraping Deskripsi Produk," *Jurnal Teknoinfo*, Vol. 14, 2020.
[11] A. I. Syahyadi, N. Afif, Ridwang, And W. Saputra, "Integrasi Data Akademik Pergururuan Tinggi Dengan Pangkalan Datadikti Menggunakan Sistem Integrasi Feeder Terbaru (Sifeeka)," *Jurnal Instek : Informatika Sains Dan Teknologi*, Vol. 8, Mar. 2023.
[12] T. J. Riasinir And Widyasari, "Pemanfaatan Framework Bootstrap Dalam Merancang Website Responsif Untuk Toko D2 Adventure," *Jurnal Enter*, Vol. 2, Aug. 2019.