Terbit online pada laman web jurnal: **https://ejournal.upm.ac.id/index.php/energy/**

Jurnal Energy

(Jurnal Ilmiah Ilmu-ilmu Teknik)

Vol. 10 No. 2 Edisi November 2020 ISSN: 2088-4591 * E-ISSN: 2962-2565

Rancang Bangun Aplikasi Web Crawling Dengan Fitur Pencarian Berita Yang Relevan Berbasis TF-IDF

Design and Build a Web Crawling Application with Relevant News Search Feature Based on TF-IDF

[1] Andika hidayatullah, [2] Imam Marzuki, S.ST., M.T, [3] Ira Aprilia, S.Pd,.M.Si
[1] Mahasiswa Program Studi Teknik Elektro, Fakultas Teknik, Universitas Panca Marga
[2,3]Dosen Program Studi Teknik Elektro, Fakultas Teknik, Universitas Panca Marga
[1]Email : andikahd7@gmail.com

Abstract

The abstract is to be in fully-justified italicized text, at the top of the paper with single column as it is here, below the author information. Use the word "Abstract" as the title, in 10-point Times, boldface type, left relative to the column, initially capitalized. The abstract is to be in 10-point, single-spaced type, and up to 200 words in length. List three to six keywords related to the articles, then continued with abstract in bahasa Indonesia.

Keywords: *abstract keywords*

Abstrak

Sebagaimana kita ketahui bersama dalam hal ini *Search Engine* (Mesin Pencari) telah banyak bermunculan dan menjadi alat memudahkan user mencari sesuatu di internet untuk menambah ilmu dan wawasan dengan luas namun dalam hal ini masih banyak user yang mengalami kesulitan dalam mengcari berita dengan *keywords* yang tidak tepat kedalam suatu web yang user tuju. Maka dari itu peneliti merancang sebuah system aplikasi *search engine* untuk website yang dikelolah merupakan web resmi yang dimiliki kampus Universitas Panca Marga menjadi bahan pembahasan bahkan dijadikan sumber data yang dikelolah dengan istilah crawling dengan mengambil data terupdate pada web tersebut. Adapun tujuan dirancangnya system web crawling ini adalah bertujuan untuk memudahkan user untuk mengakses berita terbaru pada website resmi Universitas Panca Marga dengan membatasi permasalahan hanya berita terupdate saja yang dikelolah aplikasi ini dengan cara mengcrawling pada system aplikasi ini, system aplikasi ini menggunakan database yang berisikan tentang data yang ada pada website Universitas Panca Marga dan aplikasi ini berbasis online dan harus memiliki acces internet dengan baik. Hasil pengujian aplikasi 100% berhasil dengan menggunakan pengujian blackbox testing.
Kata kunci : Crawling, Search Engine, TF-IDF, Universitas Panca Marga

1. Pendahuluan

1.1 Latar Belakang

Perkembangan teknologi web dan internet yang ada saat ini memungkinkan seseorang membuat website yang diinginkan menjadi lebih mudah. Meskipun orang tersebut memiliki pengetahuan tentang pemrograman berbasis web ataupun tidak. Cukup banyak website terbentuk tiap tahunnya (netcraft web survey).Berbagai website tersebut berusaha menjadi situs yang terkenal dan paling banyak dicari di internet terutama di search engine (mesin pencari).

Konten informasi pada saat ini menjadi kebutuhan banyak orang dalam menjalani kegiatan sehari-hari. Seperti bisa kita lihat sendiri, konten informasi berupa berita ataupun artikel- artikel yang menarik dapat menjadi salah satu kebutuhan bagi para konsumennya terhadap konten informasi yang di sajikan. Konten informasi yang disajikan juga ditunjang pada proses penyampaian konten tersebut. Proses penyampaian konten di internet yang baik merupakan tugas daripada seorang ahli teknik informatika untuk merancang dan menyajikan proses penyampaian konten yang baik melalu situs yang dibuat. Salah satu yang lebih penting ketika kita berbicara konten adalah proses mendapatkan konten itu sendiri. Proses mendapatkan konten guna memenuhi kebutuhan konten sebuah situs berita atau artikel-artikel dapat ditempuh dengan berbagai cara, salah satunya adalah dengan memanfaatkan aplikasi web crawler Aplikasi web crawler ini merupakan aplikasi yang mungkin tidak banyak digunakan dalam

DOI: https://doi.org/10.51747/energy

proses mendapatkan konten tetapi mungkin apabila dibangun dengan tujuan yang spesifik akan memiliki keuntungan yang efisien bagi yang menggunakannya sebagai proses mendapatkan konten.

Web Crawler mulai ditulis pada awal tahun 1993, pada tahun ini lahir empat web crawler pertama yaitu World Wide Web Wanderer, Jump Station, World Wide Web Worm dan RBSE Spider. Keempat web crawler ini mempunyai fungsi dasar yaitu mengumpulkan informasi dan statistik sebuah website dengan menggunakan seed URL yang kemudian akan melakukan proses download halaman situs tersebut. Sebuah web crawler (juga dikenal sebagai robot atau spider) adalah sistem untuk melakukan download sebagian besar halaman web. Web crawler digunakan untuk berbagai tujuan. Tujuan yang paling sering digunakan adalah menjadikan web crawler sebagai salah satu komponen utama dari web search engine (mesin pencari web). Selain sebagai komponen utama dari web search engine, web crawler juga bisa digunakan pada aplikasi web pengarsipan, dimana halaman web dengan skala yang besar secara berkala dikumpulkan dan diarsipkan.

Sebagian besar data dan informasi yang didapat berasal dari internet dimana informasi terus berkembang dan mengalami perubahan setiap waktu. Hal tersebut menyebabkan banyaknya data dan informasi yang tersebar di internet. Untuk memudahkan user dalam mencari data, pada umumnya digunakan search engine untuk mencari data dan informasi yang dibutuhkan. Namun untuk mengumpulkan berita dengan jumlah yang besar, metode dengan search engine bukan merupakan metode yang efektif dan efisien karena pencarian dilakukan dengan membuka link website satu persatu. Tentu hal ini akan menghabiskan waktu penggunaan.

Oleh karena itu dibutuhkan sistem pengumpulan informasi menggunakan metode crawling. Crawling merupakan cara yang digunakan untuk mengumpulkan informasi mengenai apa yang ada di halaman-halaman publik. Tujuan utamanya adalah mengumpulkan data sehingga ketika pengguna internet mengetikkan kata pencarian di website, crawling dapat dengan segera menampilkan berita yang relevan lalu menyimpannya ke dalam database sistem yang kemudian akan diklasifikasi berdasarkan tingkat pendidikan dan kategori terntentu.

Web yang akan kami crawling yaitu web pada Universitas Panca Marga probolinggo (UPM) sebagain bahan acuan bahan pokok penelitian ini, pada web UPM kami meng crawling berita judul pada halaman web upm dan mengupdate setiap berita baru yang dipost oleh web upm tersebut.

Dari hasil crawling pada web upm proses selanjutnya yaitu melakukan search engine dari beberapa data atau berita yang telah didapat Dalam konteks internet, search engine merujuk pada www atau website pada UPM. Search engine adalah sebuah program komputer yang dirancang sebagai alat bantu untuk mencari informasi di internet dengan cara mengetikkan kata kunci (keyword) yang dimaksud sehingga akan ditampilkan pada hasil pencarian yang berupa website asli yang berisi berbagai bentuk informasi seperti tulisan, Search engine memberikan pencarian content media dengan kriteria yang spesifik (berisi kata atau frasa yang ditentukan) dan memperoleh daftar file yang memenuhi kriteria tersebut. Metode TF-IDF merupakan metode yang peneliti gunakan untuk proses penggabungan dua konsep untuk perhitungan bobot, yaitu frekuensi kemunculan sebuah kata di dalam sebuah dokumen tertentu dan inverse frekuensi dokumen yang mengandung kata tersebut. Frekuensi kemunculan kata di dalam dokumen yang diberikan menunjukkan seberapa penting kata itu di dalam dokumen tersebut. Frekuensi dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut. Sehingga bobot hubungan antara sebuah kata dan sebuah dokumen akan tinggi apabila frekuensi kata tersebut tinggi di dalam dokumen dan frekuensi keseluruhan dokumen yang mengandung kata tersebut yang rendah pada kumpulan dokumen

2 Tinjauan Pustaka

2.1 Tinjauan Pustaka

Studi literatur yang penulis lakukan selanjutnya yaitu membaca beberapa penelitian-penelitian yang berkaitan dengan tugas akhir ini sebagai rujukan dan perbandingan pada metode yang digunakan serta hasil yang dicapai pada penelitian ini.

Penelitian yang dilakukan oleh sarwosri, yang berjudul “Aplikasi Web Crawler untuk web ontent pada mobile phone” Penelitian ini terfokus pada proses filtering mobile content dengan menggunakan beberapa identifikator yang bertugas menyeleksi suatu web page, hanya web page yang sesuai untuk peralatan mobile atau berisi mobile content yang akan diteruskan.

DOI: https://doi.org/10.51747/energy

Penelitian yang dilakukan oleh Agustino Halim pada tahun 2017 yang berjudul "Perancangan aplikasi web crawler untuk menghasilkan dokumen teks pada domain tertentu". Pada Penelitian ini mengahasilkan pengujian black box menggunakan metode robustness testing, aplikasi web crawler yang dibangun belum dapat menghilangkan konten iklan dari google yang terdapat pada artikel disebabkan iklan tersebut disisipkan pada awal dan akhir artikel.

Penelitian yang dilakukan oleh Astuti Dwityaning Karina, pada tahun 2009 yang berjudul "Aplikasi Performansi Web Crawling Berbasis Backlink, Breadth First Search dan Pagerank" Pengujian ini menganalisa algoritma backlink yang menunjukkan performansi terbaik dibandingkan algoritma pagerank dan BFS dalam hal melakukan crawling dan download dokumen yang memiliki tingkat kepentingan yanag tinggi.

2.2 Web *Crawling*

Studi literatur yang penulis lakukan selanjutnya yaitu membaca beberapa penelitian-penelitian yang berkaitan dengan tugas akhir ini sebagai rujukan dan perbandingan pada metode yang digunakan serta hasil yang dicapai pada penelitian ini.

Penelitian yang dilakukan oleh sarwosri, yang berjudul "Aplikasi Web Crawler untuk web ontent pada mobile phone" Penelitian ini terfokus pada proses filtering mobile content dengan menggunakan beberapa identifikator yang bertugas menyeleksi suatu web page, hanya web page yang sesuai untuk peralatan mobile atau berisi mobile content yang akan diteruskan.

Penelitian yang dilakukan oleh Agustino Halim pada tahun 2017 yang berjudul "Perancangan aplikasi web crawler untuk menghasilkan dokumen teks pada domain tertentu". Pada Penelitian ini mengahasilkan pengujian black box menggunakan metode robustness testing, aplikasi web crawler yang dibangun belum dapat menghilangkan konten iklan dari google yang terdapat pada artikel disebabkan iklan tersebut disisipkan pada awal dan akhir artikel.

Penelitian yang dilakukan oleh Astuti Dwityaning Karina, pada tahun 2009 yang berjudul "Aplikasi Performansi Web Crawling Berbasis Backlink, Breadth First Search dan Pagerank" Pengujian ini menganalisa algoritma backlink yang menunjukkan performansi terbaik dibandingkan algoritma pagerank dan BFS dalam hal melakukan crawling dan download dokumen yang memiliki tingkat kepentingan yanag tinggi.

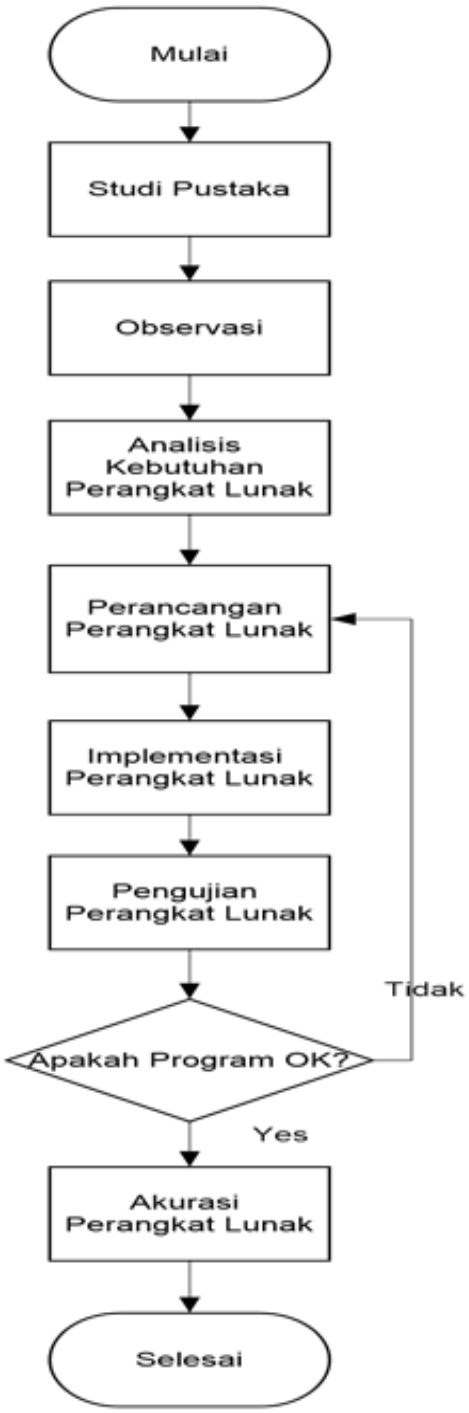
2.3 Search Engine

Search engine atau mesin pencari adalah sebuah sistem software atau sebuah aplikasi yang didesain dan disediakan oleh suatu badan komersial tertentu dan dibuat dengan tujuan untuk memudahkan pengguna internet mencari berbagai informasi di internet. Berbagai informasi ini biasanya tersimpan dalam WWW (World Wide Web), FTP (File Transfer Protocol), atau mailing list yang berada dalam sebuah server. Secara sederhananya, berbagai informasi ini nantinya akan didapatkan dari berbagai situs web, blog, atau forum yang ada di internet. Jadi, ketika pengguna mencari sebuah informasi dengan keyword atau kata kunci tertentu, search engine akan bekerja untuk mengumpulkan informasi-informasi tersebut dan menampilkan yang paling berkualitas untuk penggunanya. Setiap search engine punya kriteria dan mesin tersendiri untuk menentukan mana situs web yang berkualitas terkait dengan keyword yang dicari. Hasil pencarian yang didapat dari search engine ini nantinya akan disebut sebagai SERP atau Search Engine Result Pages.

DOI: https://doi.org/10.51747/energy

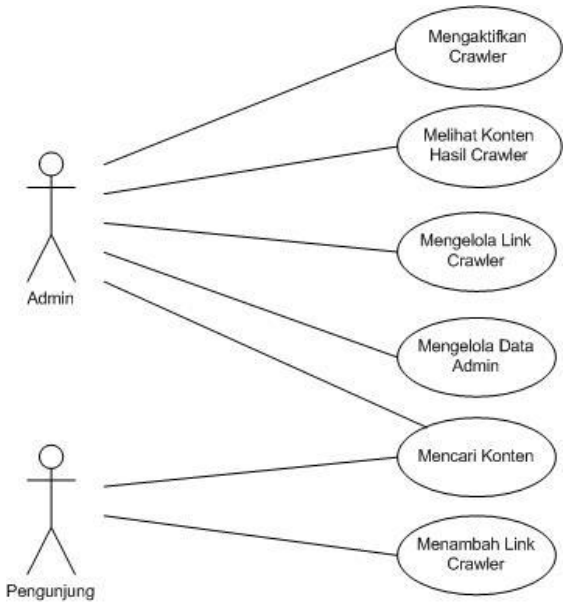
2. Metodologi

2.1 Metode Penelitian



Gambar 3.1 Diagram Alir Penelitian

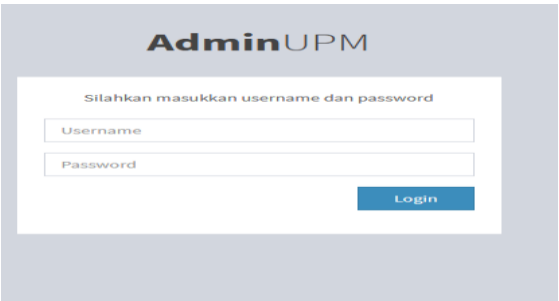
DOI: https://doi.org/10.51747/energy



Gambar 3.2. *Usecase* Rancangan Aplikasi

3. Hasil dan Pembahasan

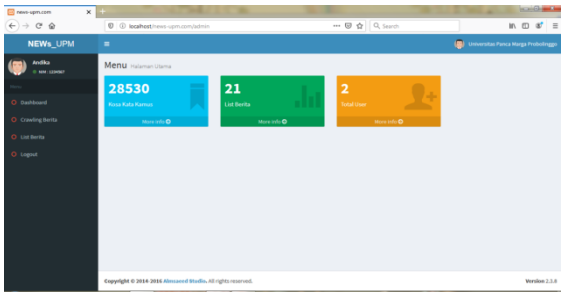
4.1 Desain Login Petugas



Gambar 4.1 Desain Menu Login

Pada gambar di atas masukkan username dan password, apabila data yang dimasukkan salah maka akan kembali ke manu login akan tetapi jika data yang dimasukkan benar maka akan masuk ke menu utama.

4.2 Desain Menu Utama

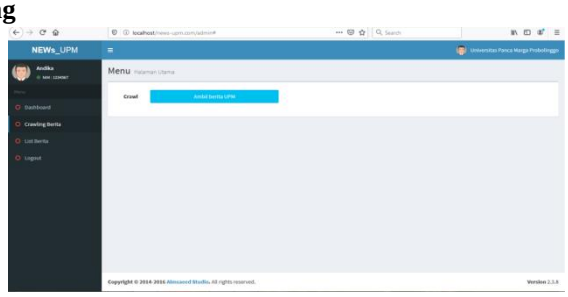


Gambar 4.2. Desain menu utama

DOI: https://doi.org/10.51747/energy

Pada gambar di atas menunjukkan tampilan menu utama di mana pada menu tersebut akan memberikan informasi – informasi yaitu kata kamus (stopword) , banyaknya berita yang telah di crawling dan terakhir infromasi data admin.

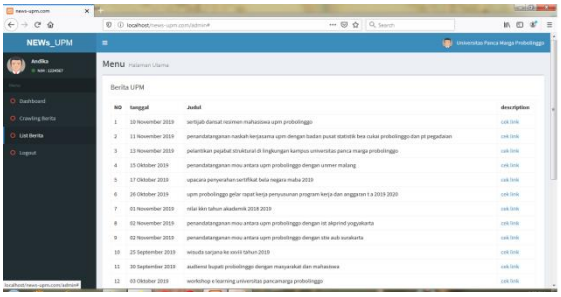
4.3 Desain Menu Crawling



Gambar 4.3. Desain Menu Crawling

Pada gambar di atas adalah form untuk *crawling* data berita pada website www.upm.ac.id yang mana untuk proses *crawling* tersebut harus menggunakan koneksi internet atau *online.*

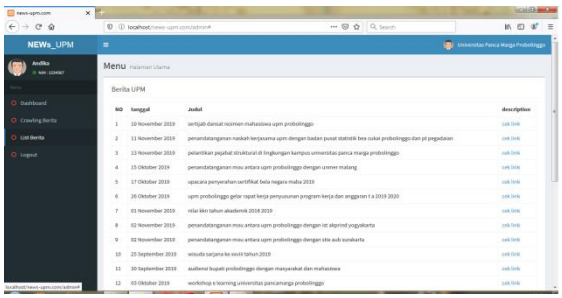
4.4 Desain Menu Berita



Gambar 4.4 Desain Menu Berita

Pada gambar di atas adalah daftar berita yang telah berhasil diambil (crawling), yang tentunya kita dapat melihatnya secara offline. Dan untuk melihat berita lengkapnya klik pada tombol “cek link” yang tentunya membutuhkan koneksi internet.

4.5 Desain Menu Pencarian Data



Gambar 4.4 Desain Menu Pencarian Data

Pada gambar di atas adalah mesin pencarian atau search engine untuk mecari berita secara akurat dan cepat seperta contoh gambar diatas mencari berita dengan kata kunci “universitas” ysng mana sistem akan menampilkan data sesuai dengan kata kunci yang dicari.

4. Kesimpulan

Kesimpulan pada penilitian ini yaitu, sistem aplikasi ini dirancang bertujuan untuk memfilter berita pada web yang kita tuju layaknya kinerja google yang bekerja sebagai search engine (mesin pencari), web crawling ini menggunakan metode TD-IDF sebagai alat deteksi kalimat dan kata atau frasa yang terdapat pada web yang kita akan proses, web yang akan di crawling merupakan web resmi milik Universitas Panca Marga namun dibatasi hanya dalam berita terbaru saja yang akan kami crwaling bertujuan untuk pengguna aplikasi ini dalam hal ini peneliti sudah memiliki ijin resmi dari pihak kampus.

DOI: https://doi.org/10.51747/energy

Saran

Adapun saran pada penelitian ini adalah sebagai berikut :

1. Mencoba dengan metode metode lainnya yang bisa dijadikan pembanding mengingat metode TF-IDF harus melakukan experiment yang lebih
2. Pada teks dengan kalimat yang panjang diperlukan perlakuan khusus atau mencari suatu metode yang dapat mengatasi teks dengan kalimat yang panjang, menimbang metode yang peniliti gunakan ini hanya menjumlah bobot kata pada suatu kalimat.

Referensi

Astuti, dkk. 2009. Analisis Performansi Web Crawler Berbasis Backlink, Breadth First Search, Dan Pagerank. Fakultas Teknik Informatika Universitas Telkom.

Dwiono, A. 2013. Mesin Pencari Cerdas dengan Web Semantik. Universitas Islam Attahiriyah Jakarta Selatan.

Himawan, dkk. 2017. Search Engine Optimization (Seo) Menggunakan Metode White Hat Seo Untuk Meningkatkan Peringkat Dan Trafik Kunjungan Website. Fakultas Ilmu Komputer Universitas Sriwijaya.

Librian, A., Kukuh, R. 2017. Sastrawi/StopWordRemover, URL: http://sastrawi.github.io.