AINET Jurnal Informatika
e-ISSN : 2686 – 1917
p-ISSN : 2657 – 0653

Peningkatan Akurasi Mesin Pencari Gambar dengan menggunakan *Machine Learning*, *Web Scrap* Dan Algoritma *Cosine Simalirity*

Adriani*[1], Ridwang[2]
[1]Program Studi Teknik Elektro Univeristas Muhammadiyah Makassar
[2]Program Studi Teknik Elektro Univeristas Muhammadiyah Makassar
e-mail: adriani@unismuh.ac.id*

Abstract

Image data search on the internet is limited to keywords in the form of image names that are entered into search engines so that the results obtained are many variants of images. With the development of technology, information retrieval and text processing are expected to help the image search process to be more specific in accordance with the desired text keywords. The web scrap process can help find more detailed information to metadata from image sources on the website. The data generated in the web scrap process is further processed using text and a cosine similarity algorithm to produce information relevant to the image being searched for. Up to 80% accurate results for general image data with image search of up to 20 images. For certain images, the accuracy is only 25% for the 20 image limit. There are 2 things that affect the assessment of image searches, namely a very large image limit and very specific keywords so that the results produced are less relevant. With an image search method like this it is expected to be able to find and download images that are truly relevant and of high quality to be used as training data in the image and video classification process.

Keyword*: Search Engines; Web Scrap; Cosine Similarity; Machine Learning*

Abstrak

Pencarian data gambar di internet terbatas pada kata kunci teks berupa nama gambar yang diinput masuk ke mesin pencari sehingga hasil yang didapatkan banyak varian gambar. Dengan berkembangnya teknologi temu kembali informasi dan pengolahan teks diharapkan dapat membantu proses pencarian gambar menjadi lebih spesifik sesuai dengan kata kunci teks yang diinginkan. Proses *web scrap* dapat membantu mencari informasi yang lebih detail sampai metadata dari sumber gambar di situs web. Data teks yang dihasilkan dalam proses *web scrap* diolah lebih lanjut menggunakan pemrosesan teks dan algoritma *cosine similarity* untuk menghasilkan informasi yang relevan dengan citra yang dicari. Hasil yang diperoleh akurasi mencapai 80% untuk data gambar umum dengan batas pencarian gambar hingga 20 gambar. Untuk gambar tertentu, akurasi hanya mencapai 25% untuk batas 20 gambar. Ada 2 hal yang mempengaruhi nilai akurasi pencarian gambar yaitu batas gambar yang sangat besar dan keyword yang sangat spesifik sehingga hasil yang dihasilkan kurang relevan. Dengan metode pencarian gambar seperti ini diharapkan dapat menemukan dan mendownload gambar yang benar-benar relevan dan berkualitas tinggi untuk dijadikan data training dalam proses klasifikasi gambar maupun video.

Kata kunci: *Mesin Pencari; Web Scrap; Cosine Similarity; Machine Learning*

1. Pendahuluan

Pengembangan sistem saat ini lebih mengarah pada *Computer Vision* seperti pengenalan wajah, pengenalan gambar atau video atau pengenalan pola tertentu. Sistem seperti ini terbukti menjadi alternatif yang dapat memfasilitasi pekerjaan di berbagai bidang. Penerapan *machine learning* pada permasalahan yang terkait dengan permasalahan tersebut sangat tepat dan efektif. Perkembangan mesin learning pada objek gambar sangat pesat, terbukti banyak aplikasi yang sedang trending memiliki dasar implementasi *machine learning*. Kemampuan dan keakuratan *machine learning* bergantung pada data pelatihan yang digunakan. Banyaknya data

pelatihan akan menambah keakuratan model saat diuji atau diterapkan. Selain jumlah data yang besar, hal-hal yang mempengaruhi kualitas data dan relevansi data dengan label pada dataset tersebut [1]. Data citra dapat diperoleh dengan mengambil gambar secara langsung dengan objek dengan kamera atau dengan mengunduh gambar di internet sesuai dengan kebutuhan data latih. Salah satu kendala untuk mendapatkan data gambar yang berkualitas dan relevan di internet adalah mesin pencari biasanya menampilkan data yang bervariasi dan gambar yang diunduh memiliki resolusi yang rendah. Maka waktu yang dibutuhkan untuk mendownload satu persatu gambar sangat banyak.

Untuk mengatasi masalah tersebut maka perlu dilakukan optimasi mesin pencari dan membuat aplikasi untuk mendownload gambar secara otomatis sesuai dengan teks *query* sehingga proses pembuatan data latih dapat berjalan dengan cepat dan akurat. Metode yang digunakan dalam penelitian ini adalah dengan menggunakan metode perbandingan data konten web sumber gambar, sehingga informasi yang diperoleh dari sumber web tersebut dapat relevan dengan teks *querys* gambar yang dicari. Proses ini membutuhkan banyak tahapan mulai dari proses web scrapping, pemrosesan teks dan kategorisasi menggunakan algoritma *cosine similarity* hingga pengurutan dan pengunduhan gambar. Dari penelitian ini diharapkan dapat menjadi solusi dalam pembuatan data latih yang membutuhkan banyak gambar untuk memudahkan pengumpulan gambar.

2. Metode Penelitian

Mesin pencari gambar untuk mengekstrak gambar yang relevan dari sejumlah gambar yang tersedia disajikan berdasarkan sistem *Hadoop*. Performa dari pendekatan yang disajikan dievaluasi menggunakan berbagai metode ekstraksi ciri citra visual seperti *Block Truncation Coding*, *Fuzzy Edge Detection* dan *Local Binary Patterns* [2].

Namun, peneliti lain berpendapat bahwa setiap tahun jutaan foto dan gambar muncul di dunia maya. Sebagian besar diunduh ke penyimpanan pribadi atau tersedia untuk umum. Dengan begitu banyak informasi, kebutuhan akan pencarian gambar yang efektif sudah matang. Dan jika alat yang sangat baik telah dibuat untuk pencarian teks, pencarian gambar tetap menjadi masalah yang belum terselesaikan. Publikasi ini bertujuan untuk mengembangkan model dalam menciptakan layanan pencarian gambar yang efektif [3].

Proses pencarian file gambar di internet lebih spesifik menggunakan bantuan metode *content based image retrieval* (CBIR) dan mesin template agar objek yang sama mudah dikenali, proses pengenalan objek menggunakan *Gradient Vector Flow Snake* (GVFS) [4].

I Sonya, prihandoko mengusulkan analisis web *scraping* terkait bencana alam dilakukan dari 3 situs media online yaitu Detikcom, Liputan6, dan VivaNews. Fokus web scraping lebih pada data tidak terstruktur di web, menjadi data yang dapat dianalisis dan disimpan. Data yang diambil dari media online berupa teks artikel dengan kata kunci yang diinputkan parameternya, kemudian diekstrak menjadi format Excel (.CSV) yang dilakukan dengan bantuan *tool Web Content Extractor* (WCE) dengan menggunakan teknik *Breadth-First* Search [5].

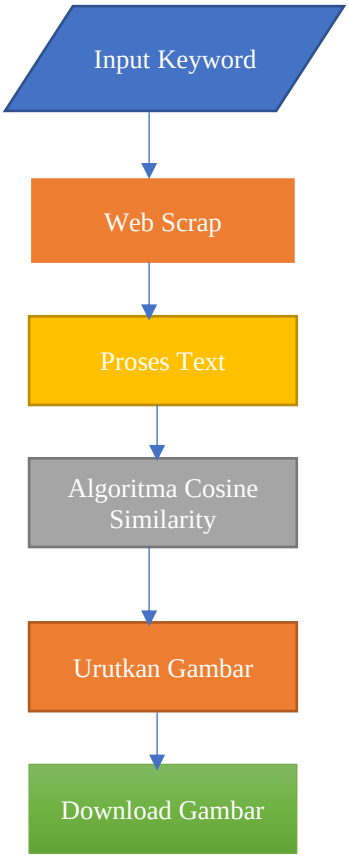
Pencarian gambar yang efektif harus didasarkan pada pendekatan data mining dalam konteks gambar, yaitu pencarian dengan kata kunci, meta data gambar, dan fitur gambar terhitung. Di sini, kata kunci dan metadata sangat berguna untuk memperbaiki area pencarian gambar dan menjadi permintaan pencarian untuk mencari informasi yang lebih rinci pada halaman web.

Efektivitas pencarian tergantung pada kompleksitas waktu komputasi fitur gambar dan komputasi teks dalam situs web. Oleh karena itu, semua perhitungan pencarian harus diminimalkan. Setiap gambar dihasilkan dari pencarian dengan menggunakan kata kunci, kemudian dilakukan proses *scraping* untuk mendapatkan informasi url sumber gambar yang nantinya akan digunakan untuk mendapatkan lebih banyak informasi dari gambar baik berupa teks maupun naratif. Hasil pengolahan teks dilanjutkan dengan mencari nilai kemiripan dengan teks *query* atau kata kunci menggunakan algoritma *cosine similarity*. Nilai dari algoritma ini dimulai dari angka 0 hingga 1, semakin dekat nilainya dengan angka 1 maka semakin tinggi tingkat kemiripan teksnya. Oleh karena itu dapat disimpulkan bahwa nilai *cosine similarity* yang tinggi pada suatu narasi di web sumber menunjukkan tingginya hubungan atau relevansi antara gambar dan teks sehingga akurasi dan efektivitas pencarian dapat ditingkatkan.

Proses selanjutnya adalah mengurutkan semua citra yang ada berdasarkan nilai *cosine similarity*-nya. Gambar dengan skor tertinggi diprioritaskan dibandingkan dengan gambar dengan skor terendah. Setelah gambar diurutkan, proses download gambar dilakukan mulai dari gambar

prioritas hingga gambar yang kurang memiliki nilai ekivalen sesuai jumlah gambar yang dibutuhkan. Algoritma cosine similarity masih memiliki beberapa kelemahan, terutama jika nilai pengolah teks berupa kata-kata bahasa Inggris atau tidak ada teks yang dihasilkan, keluaran dari algoritma ini juga akan menjadi nol. Semua ini disebabkan oleh situs sumber yang menggunakan bahasa Inggris sehingga proses pengolahan teks tidak berjalan, terutama pada proses pengukusan yang menggunakan kata-kata bahasa Indonesia. Masalah kedua adalah karena ada situs web yang tidak memiliki teks atau narasi pada tag peragraph pada halaman web sumber.

Penelitian ini memiliki beberapa tahapan proses yang dijelaskan pada gambar 1. Pertama, pengguna memasukkan teks yang akan dicari dan diunduh, kemudian teks kueri tersebut dihubungkan dengan pencarian gambar bing. Gambar yang dihasilkan menggunakan pencarian *Bing* dilanjutkan pada proses *scrapping*, dimana pada proses ini diambil URL gambar dan URL website sumber gambar. Selanjutnya dilakukan proses pengolahan teks dari halaman website sumber gambar untuk mengekstrak beberapa informasi. Proses selanjutnya adalah menghitung tingkat kemiripan teks antara input *query* dan teks website. Terakhir, unduh semua gambar dalam urutan dari nilai kemiripan tertinggi hingga terendah.



Gambar 1. Rancangan umum sistem

2.1 Data Inputan

Data inputan berupa nama gambar yang akan dicari. Teks masukan yang digunakan sebagai kata kunci pencarian disebut *query*. *Query* ini terhubung ke mesin pencari gambar *Bing* untuk melakukan pencarian gambar sesuai dengan *query* teks. Hasil pencarian gambar dari *Bing* akan diproses lebih lanjut untuk mendapatkan gambar yang lebih spesifik sesuai teks *query*.

2.2 *Scraping Web*

Scrap web adalah teknik untuk mendapatkan informasi dari situs web secara otomatis tanpa harus menyalinnya secara manual. Tujuan web *scraper* adalah untuk menemukan

informasi tertentu dan kemudian mengumpulkannya di web baru. *Scraping web* berfokus pada memperoleh data dengan cara pengambilan dan ekstraksi. Manfaat dari web *scraping* adalah informasi yang diekstraksi lebih terfokus, sehingga lebih mudah untuk mencari sesuatu. Aplikasi web *scraping* hanya berfokus pada bagaimana memperoleh data melalui pengumpulan dan ekstraksi data dengan ukuran data yang bervariasi [6].

Proses *scraping* website dari mesin pencari *Bing* dimana *Bing* adalah mesin pencari milik Microsoft. Alasan menggunakan *Bing* sebagai mesin pencari gambar adalah karena mesin pencari Google telah dibatasi pada proses *scrap* dan struktur HTML juga, sehingga sulit untuk mengikis proses gambar google. Langkah pertama dalam proses memo adalah mengambil URL gambar dan URL situs web sumber. Setelah itu, proses pengambilan HTML dilakukan pada URL situs utama untuk mengambil semua informasi dalam satu halaman HTML. Semua tag HTML dihapus dan menggabungkan semua teks terpisah.

2.3 Pengolahan Teks

Proses pengolahan teks memiliki beberapa tahapan yaitu *case folding*, *tokenizing*, *filtering* dan *stemming*. *Case folding* adalah proses mengubah semua huruf kapital menjadi huruf kecil. *Tokenizing* adalah memisahkan kalimat sesuai tujuannya menjadi kata-kata untuk diproses lebih lanjut. Sedangkan *filtering* artinya mengambil kata-kata yang penting atau berpengaruh terhadap hasil yang ingin dicapai. Proses ini dapat digunakan dengan dua cara yaitu *stoplist* dan *wordlist*. Tahap terakhir adalah *Steamming*, yaitu mencari kata-kata dasar dari kata imbuhan [7].

2.4 Persamaan *Cosine*

Cosine Similarity adalah metode yang digunakan untuk menghitung kemiripan antara dua objek. Secara umum, perhitungan metode ini didasarkan pada ukuran kesamaan ruang vektor. Metode *cosine similarity* menghitung kesamaan dalam (1) antara dua objek (misalnya D1 dan D2) yang diekspresikan dalam dua vektor menggunakan kata kunci dokumen sebagai ukuran [8].

$$similarity(q_i, d_j) = \frac{V_{q_i} \, x \, V_{d_j}}{|V_{q_i}| \, x \, |V_{d_j}|} \quad (1)$$

Hasil dari metode ini adalah antara 0 dan 1, jika tingkat kemiripan tidak ada maka bernilai 0 dan jika tingkat kesamaan mencapai 100% akan menjadi 1. jika tingkat kesamaan di bawah 100% maka nilainya di kisaran antara 0 dan 1.

2.5 *Bubble Short*

Algoritma adalah struktur yang diterapkan pada bahasa atau pemrograman komputer dengan tujuan membantu menyelesaikan masalah dimana akan terdapat data sebagai masukan dan keluaran sebagai hasil dari proses yang dilakukan. Algoritma pengurutan memiliki kelebihan dan kekurangan masing-masing, tidak semua algoritma dapat digunakan untuk mengurutkan data. Algoritma *bubble sort* adalah salah satu dari beberapa jenis pengurutan yang digunakan untuk mengurutkan data. Cara kerja algoritma ini adalah dengan mengulang suatu proses, kemudian membuat perbandingan pada masing-masing elemen larik dan mengubah posisi jika urutannya benar. Perbandingan setiap elemen dari larik ini akan berlanjut hingga kondisi yang ditentukan terpenuhi. Jenis algoritma ini termasuk dalam algoritma sortir perbandingan, karena melakukan perbandingan dalam operasi antara elemen *array* yang disediakan [9].

Algoritma *bubble short* dibagi menjadi dua tahapan. Dimana pada tahapan pertama membandingkan larik y[1] dan y[2]. Sedangkan pada tahapan kedua membangdingkan y[n-2] dengan y[n-1].

Tahapan pertama :

a) Membandingkan larik y [1] dengan larik y [2], kemudian menyusunnya kembali berdasarkan urutan yang telah disesuaikan, sehingga y [1] <y [2].
b) Perbandingan larik y [2] dengan larik y [n], kemudian susun kembali berdasarkan urutan yang telah disesuaikan, sehingga y [2] <y [n].
c) Membandingkan larik y [n-1] dengan larik y [n], kemudian menyusunnya kembali berdasarkan urutan yang telah disesuaikan, sehingga larik y [n-1] <y [n], setelah (n-1) dibandingkan, y [n] akan menjadi elemen array terurut terbesar atau terkecil.

Tahapan kedua :

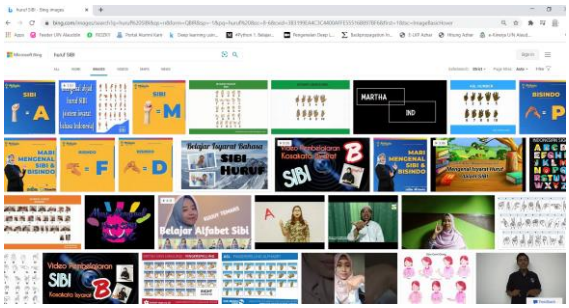
a) Ulangi perbandingan bagian kedua sampai Anda telah membandingkan dan kemungkinan susunan untuk y [n-2], y [n-1]
b) Setelah perbandingan larik elemen (n-2), (n-1) akan menjadi elemen terbesar kedua
c) Dan lanjutkan langkah selanjutnya Langkah ke (n-1)
d) Bandingkan y [1] dengan y [2] lalu susun kembali sehingga muncul barisan y [1] <y [2]. Setelah elemen array mengambil langkah (n-1), elemen array akan disusun dalam urutan naik atau turun sesuai dengan kondisi yang telah ditentukan.
e) Dan lanjutkan langkah selanjutnya hingga proses terakhir selesai.

2.6 Unduh Gambar

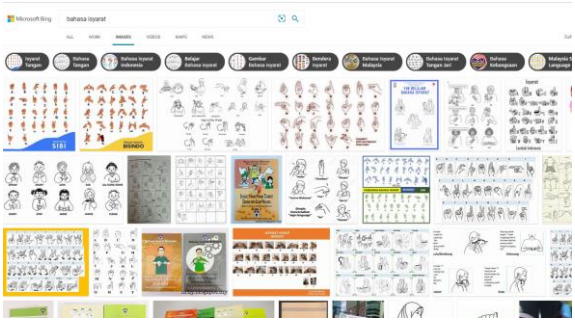
Proses terakhir adalah mengurutkan data citra berdasarkan nilai kesamaan tertinggi hingga nilai terendah. Pengurutan data menggunakan teknik *bubble short sorting*, dimana teknik ini melakukan pengurutan secara detail dan tepat dari nilai awal hingga nilai akhir. Setelah data citra tersusun rapi maka proses pengunduhan citra dilakukan ke penyimpanan komputer lokal sesuai jumlah citra yang diinput.

3. Hasil dan diskusi

Percobaan dilakukan dengan menggunakan 2 kata kunci yang berbeda, yaitu kata kunci yang bersifat umum dan kata kunci yang bersifat spesifik.



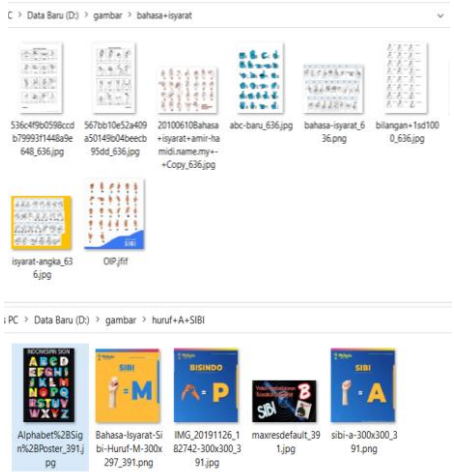
Gambar 2. Keyword spesifik



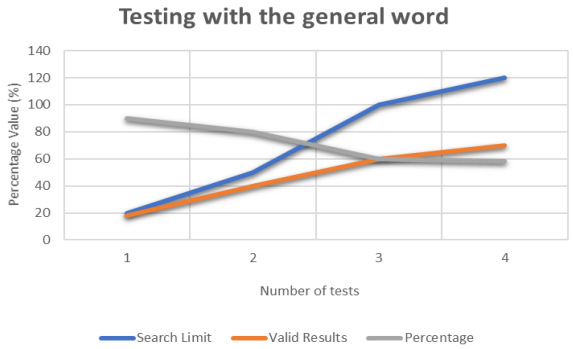
Gambar 3. Keyword umum

Pada gambar diatas sudah terlihat bahwa pencarian dengan menggunakan *keyword* yang berbeda menampilkan hasil yang signifikan perbedaanya. Semakin umum kata kunci yang digunakan maka hasil yang akan ditampilkan mesin pencari begitupun sebaliknya, semakin spesifik kata kunci yang digunakan maka hasil yang dihasilkan semakin sedikit. Oleh karena itu mengoptimalkan pencarian dan download gambar yang relevan dengan *keyword* maka dibutuhkan algoritma dan *machine learning* untuk mengenali gambar yang di download apakah sesuai dengan *keyword*. Gambar hasil download memiliki jumlah lebih sedikit dibanding dengan jumlah gambar yang dihasilkan mesin pencari karena untuk melakukan proses downloading gambar di dahului oleh proses filter menggunakan *web scrap*, *text processing* dan algoritma *cosine similarity*. Jadi pada intinya, gambar yang di download betul – betul memiliki relevansi yang besar dengan kata kunci pencarian [10].

AINET Jurnal Informatika
e-ISSN : 2686 – 1917
p-ISSN : 2657 – 0653

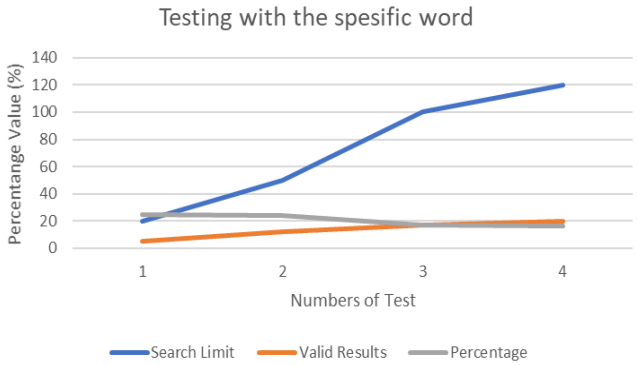


Gambar 4. Gambar yang berhasil di download



Gambar 5. Grafik testing menggunakan kata kunci umum

Berdasarkan grafik pada Gambar 5, dapat disimpulkan bahwa tingkat akurasi meningkat jika jumlah gambar yang dicari sedikit dengan kata-kata umum. Semakin banyak gambar yang dicari maka semakin mengurangi nilai keakuratannya karena disebabkan oleh terbatasnya jumlah gambar yang ditemukan oleh mesin pencari dalam satu proses pencarian.



Gambar 6. Grafik testing menggunakan kata kunci khusus

Untuk data kedua yang menggunakan kata / *query* yang tidak umum, maka gambar hasil pencarian yang valid juga berkurang karena gambar hasil *query* juga dikurangi oleh mesin pencari. Kesimpulannya, semakin banyak batasan atau jumlah gambar dalam hasil pencarian, maka gambar yang dihasilkan akan semakin kurang akurat. Nilai akurasi tertinggi hanya sampai 25% dengan batas 20 gambar.

4. Kesimpulan

Pencarian dan download gambar di internet merupakan salah satu alternatif atau solusi untuk mendapatkan data latih dalam jumlah besar dalam proses klasifikasi atau pengelompokan gambar, oleh karena itu perlu dikembangkan teknik pencarian gambar agar hasilnya sesuai dengan kebutuhan data latih baik dari segi kuantitas, relevansi, dan kualitas gambar. Penelitian yang telah dilakukan adalah melakukan peningkatan kualitas pencarian dengan menggunakan metadata dari website sumber gambar sebagai bahan referensi yang mengacu pada gambar tersebut. Scraping web dan pengolahan teks merupakan inti dari proses untuk mendapatkan data teks dari website yang digunakan sebagai input data pada algoritma cosine similarity. Hasil yang didapat dari proses ini hanya mencapai akurasi 50%. Melakukan optimasi pencarian gambar dengan web scrapping kurang akurat karena terdapat beberapa halaman web sumber gambar yang tidak memiliki narasi sehingga proses klasifikasi teks tidak berjalan dengan baik dan harus merujuk kepada web yang mempunyai template yang sama.

5. Notasi

V_{q_i} : Query Vector
V_{d_j} : Vektor Dokumen

Referensi

[1] M. Arsenovic, S. Sladojevic, A. Anderla, D. Stefanovic, and B. Lalic, “Deep learning powered automated tool for generating image based datasets,” in *2017 IEEE 14th International Scientific Conference on Informatics*, 2017, pp. 13–17.

[2] D. Uttarwar, A. Agarwal, R. Kadiwar, and V. D. Katkar, “Distributed content based image search engine using hadoop framework,” in *2017 International Conference on Communication and Signal Processing (ICCSP)*, 2017, pp. 1706–1710.

[3] K. Smelyakov, D. Sandrkin, I. Ruban, M. Vitalii, and Y. Romanenkov, “Search by image. New search engine service model,” in *2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T)*, 2018, pp. 181–186.

[4] I. Wahdaniah and H. Harlinda, “Berdasarkan Fitur Bentuk Berbasis Web Menggunakan Metode Gradien Vektor Flow Snake,” *Ilk. J. Ilm.*, vol. 9, no. 1, pp. 49–56, 2017.

[5] I. P. Sonya and P. Prihandoko, “Analisis Web Scraping untuk Data Bencana Alam dengan Menggunakan Teknik Breadth-First Search Terhadap 3 Media Online,” *J. Ilm. Inform. Komput.*, vol. 21, no. 3, 2017.

[6] D. D. A. Yani, H. S. Pratiwi, and H. Muhardi, “Implementasi Web Scraping untuk Pengambilan Data pada Situs Marketplace,” *JUSTIN (Jurnal Sist. dan Teknol. Informasi)*, vol. 7, no. 4, pp. 257–262, 2019.

[7] R. RIDWANG, N. U. R. AFIF, and A. I. SYAHYADI, “SISTEM PAKAR PENDETEKSI PENYAKIT PADA BALITA MENGGUNAKAN METODE COSINE SIMILARITY DAN ALGORTIMA NAZIEF ADRIANI,” *J. INSTEK (Informatika Sains dan Teknol.*, vol. 5, no. 1, pp. 57–66, 2020.

[8] O. Nurdiana, J. Jumadi, and D. Nursantika, “Perbandingan metode Cosine Similarity dengan metode Jaccard Similarity pada aplikasi pencarian terjemah Al-Qur’an dalam Bahasa Indonesia,” *J. Online Inform.*, vol. 1, no. 1, pp. 59–63, 2016.

[9] I. Gunawan, S. Sumarno, and H. S. Tambunan, “Penggunaan Algoritma Sorting Bubble Sort Untuk Penentuan Nilai Prestasi Siswa,” *Sist. J. Sist. Inf.*, vol. 8, no. 2, pp. 296–304, 2019.

[10] A. A. Ilham and I. Nurtanio, “Image search optimization with web scraping, text processing and cosine similarity algorithms,” in *2020 IEEE International Conference on Communication, Networks and Satellite (Comnetsat)*, 2020, pp. 346–350.