SMARTICS Journal, Vol.5 No. 1 2019. p22-26
ISSN online: 2476-9754, ISSN print: 2623-0429
DOI : https://doi.org/10.21067/smartics.v5i1.2848

Deteksi Plagiarisme pada Artikel Jurnal Menggunakan Metode *Cosine Similarity*

Rito Putriwana Pratama[1], Muhammad Faisal[2], Ajib Hanani[3]
[1,2,3]Teknik Informatika, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia
email : [1,2,3]ritoputri25@gmail.com, akhifai@gmail.com, ajib@uin-malang.ac.id

***Abstract*— Plagiarism is an act of taking ideas, taking research results, acquiring research results, and summarizing a writing without mentioning the source. The method of cosine similarity is one of the methods that can be used to calculate similarity among articles. The system stages to produce similarities among articles are comparing the journal article uploaded in the repository obtained from the results of data grabbing of DOAJ. In the calculation method performed will get the percentage of similarity among the articles. After that it will be recalculated to find the similarity of journal article among publishers in the repository. Based on the trial scenario performed by calculating the number of relevant articles divided by the number of articles then multiplied by 100%, it would be obtained the recall value on the Application of Plagiarism Detection Using the Cosine Similarity Method was 13%. While to obtain precision value, a test scenario was done by calculating the number of relevant articles divided by the number of relevant articles in the search then multiplied by 100% and it was obtained the result of 8%.**

***Index Terms*— *cosine similarity*; journal article; detection; plagiarism.**

***Abstrak*— Plagiarisme merupakan tindakan mengambil gagasan, mengambil hasil riset, mengakuisisi hasil riset, dan meringkas suatu tulisan tanpa menyebutkan sumbernya. Metode *cosine similarity* merupakan salah satu metode yang dapat digunakan untuk menghitung nilai kemiripan antar artikel. Tahapan yang dilakukan sistem untuk menghasilkan nilai kemiripan antar artikel yaitu dengan membandingkan artikel jurnal yang di upload di repository yang diperoleh dari hasil grabbing data DOAJ. Dalam perhitungan metode yang dilakukan akan diperoleh presentase nilai kemiripan antar artikel. Setelah itu akan dihitung kembali untuk mencari nilai kemiripan artikel jurnal antar publisher yang ada di repository. Berdasarkan skenario uji coba yang dilakukan dengan menghitung jumlah atikel relevan terambil dibagi dengan jumlah artikel dikali 100%, diperoleh nilai *recall* pada Aplikasi Deteksi Plagiarisme Menggunakan Metode *Cosine Similarity* yaitu 13%. Sedangkan untuk memperoleh nilai *precision* dilakukan skenario pengujian dengan menghitung jumlah artikel relevan terambil dibagi dengan jumlah dokumen relevan dalam pencarian kemudian dikali 100% diperoleh hasil 8%.**

***Kata Kunci*— *cosine similarity*; artikel jurnal ; deteksi; plagiarisme.**

I. PENDAHULUAN

Pesatnya perkembangan dunia teknologi saat ini merubah gaya hidup manusia menjadi serba digital. Teknologi merupakan kebutuhan yang tidak dapat dilepaskan dari kehidupan manusia sehari-hari di berbagai bidang. Hal ini menjadikan manusia hidup di era digital dengan berbagai dampak positif dan negatifnya. Bidang pendidikan merupakan salah satu yang menjadi dampak dari adanya perkembangan teknologi, seperti dokumen digital.

Dokumen digital merupakan dampak dari adanya perkembangan dunia teknologi di era digital seperti saat ini. Jurnal *online* merupakan dokumen digital yang sangat dibutuhkan masyarakat dalam semua bidang. Jurnal sendiri merupakan bagian dari jenis terbitan berseri yang ada di perpustakaan. Jurnal berisi koleksi dan terbitan atau transmisi mengenai berita dan hasil-hasil penelitian mengenai media. Jurnal terdapat dua format yaitu tercetak dan digital. Jurnal *online* merupakan versi digital dari jurnal cetak yang biasa ditemui di perpustakaan. Jurnal *online* tersedia melalui *emai, web* atau akses internet. Kesamaan jurnal cetak dan *online* yaitu sama sama dalam terbitan berseri, sedangkan perbedaannya terdapat pada bahan baku yang digunakan, jika jurnal cetak berbahan baku kertas sedangkan jurnal *online* tanpa dicetak dapat langsung dibaca secara *online*.

Keuntungan tersedianya jurnal *online* yaitu mudah dibaca dimana saja tanpa membawa kertas, sedangkan dampak negatif dari jurnal *online* yaitu dapat dijiplak atau perilaku plagiat. Terdapat banyak bentuk plagiat, salah satu yang sering dilakukan yaitu *copy-paste-edit* suatu artikel jurnal *online*. Plagiat dapat ditemukan dalam bentuk kutipan pada sebuah dokumen [1].

Mendeteksi plagiat dapat dilakukan dengan cara manual meski tidak efektif karena harus memeriksa sebuah artikel dengan ribuan artikel lainnya dan menafsirkan gaya penulisannya. Cara mudah untuk mendeteksi plagiat yaitu dengan penggunaan search engine atau mesin pencarian dengan memasukkan kata kunci tema artikel dan membiarkan mesin pencarian menemukan artikel yang dijiplak [1]. *Search engine* sendiri merupakan program komputer yang dapat membantu pengguna untuk menemukan informasi yang relevan dengan kebutuhan. Masukan dari mesin pencarian yaitu berupa kata kunci yang dibutuhkan. Dengan waktu yang relatif singkat, mesin pencarian akan memberi hasil berupa artikel yang relevan dengan kata kunci yang dimasukkan. Hal berikut sangat berguna apabila plagiat dilakukan pada seluruh dokumen, tetapi kurang efektif apabila plagiat dilakukan hanya pada sebagian artikel dan menggabungkan beberapa pecahan artikel lainnya.

Untuk dapat menjalankan *search engine* pendeteksi plagiarisme, pertama yang dibutuhkan yaitu sebuah cara untuk mengambil halaman web secara otomatis agar diperoleh dokumen jurnal yang *up to date*. Dalam hal ini, peneliti menggunakan metode web crawler karena dianggap mampu menjelajahi serta mengambil halaman web secara rekursif dan otomatis dengan mengikuti *hyperlink* yang tersedia kemudian mengambil URL yang diberikan agar dapat mengunduh dan mengambil link dari halaman web lain [2].

Cara yang dilakukan untuk mengambil (*filtering*) konten jurnal sebelum dicari kemiripannya yaitu dengan menggunakan pdf *extractor*. Cara ini dianggap mampu menjelajahi dan mengambil semua konten jurnal dalam bentuk pdf. Konten jurnal yang diambil diantaranya berupa metada data serta keseluruhan isi artikel pada jurnal. Selain menggunakan pdf *extractor* dalam *filtering content*, dibutuhkan metode lain untuk proses pencarian serta menilai tingkat kemiripan sebuah jurnal *online*. Salah satu metode yang digunakan peneliti dalam mesin pencarian yaitu metode *Cosine Similarity*. *Cosine similarity* yaitu metode perhitungan antara dua buah dokumen yang bertujuan untuk mengetahui tingkat kemiripan antar dokumen tersebut [3]. Perhitungan metode c*osine similarity* didasarkan pada dua buah vektor yang memiliki kemiripan jumlah kata pada dua dokumen yang dibandingkan. Peneliti menggunakan metode *Cosine Similarity* karena memiliki nilai keakuratan lebih tinggi dibandingkan dengan metode Jaccard Similarity. Hal tersebut dikarenakan metode *Cosine Similarity* mempunyai konsep normalisasi panjang vektor data dengan membandingkan N-gram yang sejajar satu sama lain dari 2 pembanding [4].

II. KAJIAN PUSTAKA

Dalam Kamus Besar Bahasa Indonesia *online* dikatakan bahwa plagiat adalah pengambilan karangan (pendapat dan sebagainya) orang lain dan menjadikannya seolah-olah karangan (pendapat dan sebagainya) sendiri, misalnya menerbitkan karya tulis orang lain atas nama dirinya sendiri. Tertulis pula dalam Peraturan Menteri Pendidikan RI Nomor 17 Tahun 2010 bahwa plagiat adalah perbuatan sengaja atau tidak sengaja dalam memperoleh kredit atau nilai untuk suatu karya ilmiah, dengan mengutip sebagian atau seluruh karya dan atau karya ilmiah pihak lain yang diakui sebagai karya ilmiahnya, tanpa menyatakan sumber secara tepat [5].

Web Crawler merupakan program yang dapat menelusuri halaman web secara rekursif dan otomatis dengan mengikuti *hyperlink* yang telah disediakan. Pada penelitian yang sebelumnya dilakukan oleh Kumar dkk dikatakan bahwa *web crawler* yang digunakan fokus berbasis *query* yang dapat dijelajah dengan cepat. *Web Crawler* ini dinilai lebih efisien daripada *crawler* BFS sebelumnya karena dapat mengambil halaman web bersama dengan tag meta untuk menentukan relevansi halaman web [6]. Untuk menemukan halaman web yang terkait dengan topik di internet, cukup dengan memasukkan satu set kata kunci kemudian memeriksa setiap halamannya. Hasilnya akan diketahui judul, deskripsi, konten dll. *Web Crawler* hanya dapat mengikuti link *hypertext* pada internet dengan menyediakan *crawler* sebagai sarana untuk menentukan relevansi serta menemukan rute terbaik untuk menjelajahi halaman web [7].

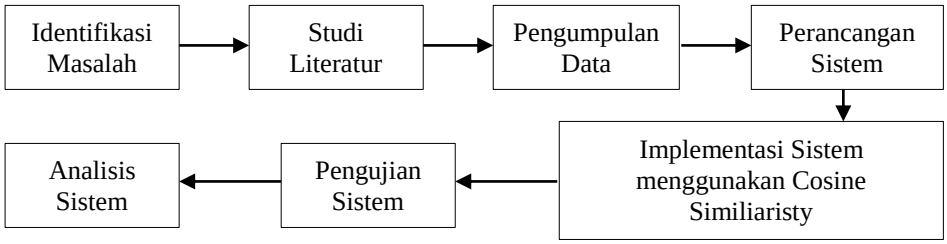
Menghitung nilai kemiripan suatu dokumen menggunakan metode *cosine similarity* yaitu dengan cara membandingkan antara vektor *query* dengan vektor dokumen. Hasil dari perbandingan tersebut berupa sudut *cosinus x* dengan 0 sebagai nilai terkecil yang berarti tidak terdapat kemiripan dan 1 sebagai nilai terbesar yang berarti memiliki nilai kemiripan yang besar. [8].

III. METODE PENELITIAN

Bab metodologi peneltian memaparkan tahap-tahap kegiatan dalam melaksanakan penelitian. Peneltian ini mengambil judul Aplikasi Deteksi Plagiarisme Menggunakan Metode *Cosine Similarity*.

A. Prosedur Penelitian

Adapun prosedur penelitian pada penelitian ini direpresentasikan ke dalam diagram pada gambar 1.



Gambar 1 Prosedur Penelitian

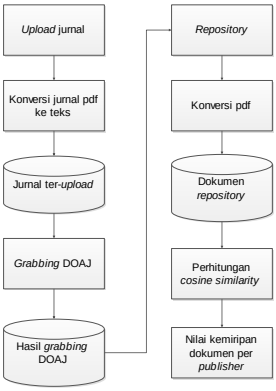
Penelitian dimulai dari identifikasi masalah dengan menentukan pertanyaan penelitian. Tahap selanjutnya yaitu studi literatur yaitu mengumpulkan teori-teori yang mendukung penelitian, seperti melakukan proses *grabbing* dan *Cosine Similarity*. Dilanjutkan pada tahap pengumpulan data, dalam penelitian ini data yang dibutuhkan adalah jurnal untuk dijadikan dokumen repository. Tahap selanjutnya adalah perancangan sistem yaitu untuk memahami alur sistem yang akan dibuat dan mengimplementasikan *web crawler* pada pengambilan konten di setiap jurnal serta metode *cosine similarity* untuk menghitung kemiripan teks dalam proses deteksi plagiarisme. Setelah melakukan perhitungan nilai kemiripan dokumen, akan dilanjutkan tahap pengujian sistem dengan memperhatikan kebenaran sistem dengan metode yang telah digunakan. Dari semua tahap diatas, dapat dilanjutkan dengan melakukan analisis dari perhitungan metode serta sistem yang telah dibuat.

B. Pengumpulan Data

- Jurnal *Online*. Obyek yang digunakan unutk penelitian yaitu jurnal *online*. Tahap pertama yaitu mengumpulkan atau mencari referensi jurnal *online* di berbagai *website* atau *blog.*
- Konversi PDF digunakan untuk mengubah file pdf menjadi teks agar dapat dihitung nilai kemiripan dengan dokumen repository yang terdapat di database.
- Isi jurnal diperoleh semua isi dokumen jurnal berupa teks yang akan dilakukan perhitungan kemiripannya.
- Database. Tahap akhir, memasukkan konten yang diperoleh ke database yang telah dibuat. Data dalam database inilah yang nantinya akan diproses pada aplikasi deteksi plagiarisme artikel jurnal.

C. Perancangan Sistem

Pada tahap ini terdapat desain rancangan sistem yang akan memberikan gambaran yang harus dikerjakan serta bagaimana sistem memecahkan masalah deteksi plagiarisme.

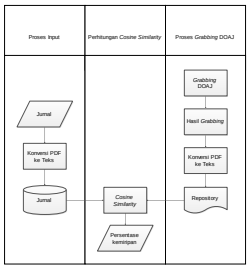


Gambar 2 Perancangan Sistem

Yang pertama yaitu proses upload jurnal. Pada proses ini, mahasiswa akan meng-upload jurnal dalam bentuk pdf untuk di cek nilai kemiripannya dengan dokumen repository. Jurnal yang di upload terlebih dahulu akan di konversi kedalam bentuk teks dan hasil konversi masuk kedalam database. Untuk mencari repository, maka proses selanjutnya yaitu grabbing jurnal yang ada di Directory of Open Access Journal (DOAJ). Grabbing bisa dilakukan sesuai dengan keinginan dengan memasukkan tema dan jumlah jurnal yang akan di ambil. Semua hasil grabbing DOAJ akan masuk pada database yang selanjutnya akan di proses kembali untuk dijadikan dokumen repository dengan mengkonversi jurnal dari bentuk pdf ke bentuk teks untuk dilakukan perhitungan algoritma cosine similarity. Perhitungan algoritma dilakukan untuk membandingkan jurnal yang di upload dengan jurnal yang terdapat pada dokumen repository.

D. Desain Sistem

Desain sistem bertujuan untuk memberikan gambaran yang jelas dan rancang bangun yang lengkap kepada pemakai sistem.



Gambar 3 Desain Sistem

- Input Jurnal yang akan di deteksi kemiripannya dengan dokumen repository yang telah tersedia di database.
- Konversi pdf ke teks agar bisa dihitung nilai kemiripannya.
- *Grabbing* DOAJ untuk menjelajah dan mengambil halaman web yang dibutuhkan dengan menelusuri *hyperlink* yang terdapat di DOAJ.
- Hitung nilai kemiripan untuk mendapatkan persentase nilai kemiripan antara dokumen jurnal dengan dokumen repository yang telah didapat.

IV. HASIL DAN PEMBAHASAN

A. Perhitungan

Pada pengujian sistem diambil beberapa dokumen uji yang akan dihitung nilai kemiripannya dengan dokumen pembanding yang telah didapat dari hasil grabbing DOAJ. nilai persentase menunjukkan hasil kemiripan antar dokumen yang dihitung menggunakan rumus cosine similarity, yaitu :

$$similarity(d_j, q) = \frac{\sum_{i=l}^{t}(w_{ij}.w_{iq})}{\sqrt{\sum_{i=l}^{t} w_{ij}^2 . \sum_{i=l}^{t} w_{iq}^2}} \quad \text{….....(1)}$$

Perhitungan *Recall*

Dari proses pencarian, diperoleh hasil 12 dokumen jurnal yang sesuai dengan nama file yang diinputkan pada proses pencarian dalam tabel dokumen repository. Dari data dokumen yang diperoleh melalui proses pencarian file ‘data mining’, dapat dilakukan perhitungan *recall* untuk mengetahui tingkat akurasi suatu sistem. Perhitungan nilai *recall* dilakukan menggunakan rumus seperti dibawah ini.

$$Recall = \frac{\sum dokumen\ relevan\ yang\ terambil}{\sum dokumen\ relevan\ dalam\ database} \text{ x } 100\% \quad \text{……..(2)}$$

$$= \frac{12}{91} \times 100\%$$

$$= 13\%$$

Perhitungan *recall* merupakan parameter untuk mengukur tingkat akurasi suatu sistem berdasarkan dokumen relevan yang terdapat dalam databse. Pada sistem ini, diperoleh nilai *recall* yaitu 13%.

Perhitungan *Precision*

Jumlah dokumen relevan yang terambil pada saat melakukan proses pencarian file ‘data mining’ yaitu 12 dokumen jurnal. Sedangkan untuk jumlah dokumen yang relevan dalam pencarian diperoleh 1 dokumen jurnal yang sesuai dengan nama file pada proses pencarian ‘data mining’ dalam dokumen repository dengan nilai *cosine similarity* diatas 0% dari total jurnal dalam pencarian sejumlah 12 file. Dari data jumlah dokumen jurnal yang diperoleh dari proses pencarian fil e’data mining’ seperti dijelaskan diatas, dapat dilakukan perhitungan nilai *precision* dengan menggunakan rumus seperti dibawah ini.

$$Precision = \frac{\sum dokumen\ relevan\ yang\ terambil}{\sum dokumen\ relevan\ dalam\ pencarian} \text{ x } 100\% \quad \text{.....(3)}$$

$$= \frac{1}{12} \times 100\%$$

$$= 8\%$$

Perhitungan *precision* merupakan parameter untuk mengukur tingkat akurasi sebuah sistem berdasarkan dokumen yang relevan pada pencarian yang dilakukan. Pada sistem ini, diperoleh nilai *precision* yaitu 8%.

B. Perangkat

Implementasi merupakan tahap penerapan sistem yang akan dilakukan jika sistem disetujui sebagai program yang telah dibuat pada tahap perancangan. Selain itu, implementasi sistem merupakan sebuah

proses pembuatan dan penerapan sistem secara utuh baik dari sisi perangkat keras maupun perangkat lunaknya.

C. Implementasi Perangkat Lunak

Untuk mendukung aplikasi agar berjalan dengan optimal, maka dibutuhkan perankat lunak pengolahan data, adapun perangkat lunak yang digunakan untuk mendukung pembuatan aplikasi itu adalah sebagai berikut :

1. PHP
2. MySQL
3. Server Apache

Selain perangkat lunak pengolahan data seperti yang diatas, pembuatan aplikasi juga membutuhkan pihak ketiga atau biasa disebut dengan Third Party. Third Party yang digunakan adalah sebagai berikut :

1. class.pdf2text.php
2. phpMaker

D. Implementasi Perangkat Keras

Perangkat keras yang dibutuhkan berdasarkan kebutuhan minimal yang harus dipenuhi adalah sebagai berikut :

1. Processor intel inside core i5
2. Memory (RAM) 4GB
3. Hardisk 500GB
4. Mouse, printer

E. Implementasi *Interface*

Interface merupakan tampilan yang dapat melakukan interaksi antara pengguna dengan sistem, dimana interface dapat menerima informasi dari pengguna dan memberikan informasi kepada pengguna yang bertujuan untuk menginput pengetahuan baru ke dalam basis pengetahuan sistem pakar, menampilkan penjelasan sistem dan memberikan panduan pemakaian sistem secara menyeluruh sehingga dapat dipahami oleh pengguna.

V. PENUTUP

Dari hasil penelitian dan pembahasan tentang aplikasi deteksi plagiarisme menggunakan metode *cosine similarity*, dapat diambil kesimpulan nilai akurasi sistem dengan melakukan perhitungan *recall* dan *precision* dari perhitungan *cosine* dengan mengambil data dan membandingkan dengan repository yang telah ada. Nilai *recall* untuk kasus ini yaitu 13%, diperoleh dari jumlah dokumen relevan yang terambil dibagi dengan jumlah dokumen yang ada dalam database dikali 100%. Sedangkan nilai *precision* yaitu 8%, diperoleh dari jumlah dokumen relevan yang terambil dibagi dengan jumlah dokumen relevan yang ada dalam pencarian dikali 100%.

DAFTAR PUSTAKA

[1] Firdaus, Hari Bagus. 2003. "Algoritma Rabin-Karp." Jurnal Ilmu Komputer dan Teknologi Informasi III No. 2: 1–5.
[2] Zuliarso, Eri. 2010. "Aplikasi Web Crawler Berdasarkan Breadth First Search Dan Back-Link." Fakultas Teknologi Informasi, Universitas Stikubank Semarang XV(1): 52–56.
[3] Sugiyamta. 2015. "Sistem Deteksi Kemiripan Dokumen Dengan Algoritma Cosine Similarity Dan Single Pass Clustering." Dinamika Informatika 7(2): 7.
[4] Nurdiana, Ogie, Jumadi, and Dian Nursantika. 2016. "Perbandingan Metode Cosine Similarity Dengan Metode Jaccard Similarity Pada Aplikasi Pencarian Terjemah Al-Qur'an Dalam Bahasa Indonesia." Jurnal Online Informatika (JOIN) 1(1): 59–63.
[5] Santoso, Hari. 2015. "Pencegahan Dan Penaggulangan Plagiarisme Dalam Penulisan Karya Ilmiah Di Lingkungan Perpustakaan Perguruan Tinggi Oleh : Drs. Hari Santoso, S.Sos. 1." Perpustakaan UM Malang (1): 1–23.
[6] Kumar, Manish, Ankit Bindal, Robin Gautam, and Rajesh Bhatia. 2018. "Keyword Query Based Focused Web Crawler." Procedia Computer Science 125: 584–90. http://linkinghub.elsevier.com/retrieve/pii/S1877050917328399.
[7] Rungsawang, Arnon, and Niran Angkawattanawit. 2005. "Learnable Topic-Specific Web Crawler." Journal of Network and Computer Applications 28(2): 97–114.
[8] Pahlevi, Irfan, Moch Arief Bijaksana, and M Ir Tech. "Perhitungan Kemiripan Dokumen Bahasa Indonesia Menggunakan Metode Cosine Similarity (Studi Kasus : Abstrak Tugas Akhir Fakultas Informatika Universitas Telkom)."