

BAB I

PENDAHULUAN

1.1 Latar Belakang

Sebuah Informasi adalah kebutuhan yang berguna untuk segala pihak dari pemerintahan maupun rakyat sipil atau masyarakat, bahkan informasi memiliki kekuatan tersendiri dan dapat mempengaruhi keadaan dari suatu tempat maupun kejadian. Informasi bahkan bisa didapatkan secara mudah di internet yang dapat diakses sepanjang waktu ketika diperlukan. Dokumen merupakan salah satu media yang menyajikan informasi bagi yang memerlukan dokumen tersebut. Kini banyaknya dokumen yang tersebar di internet memungkinkan pencarian dokumen menjadi mudah. Tetapi situs-situs tersebut tentu saja belum memberikan dokumen yang valid dan cenderung memberikan informasi yang sama atau plagiarisme, bahkan tanpa mengutip dan mengakuisisi ide asing menjadi milik mereka tanpa sepengetahuan pemilik asli ide tersebut (plagiator) [1].

Plagiarisme sebuah tindakan dimana seseorang menjiplak suatu pekerjaan, karya maupun suatu dokumen secara tidak sengaja/disengaja agar mendapatkan keuntungan dari *plagiarism* tersebut, bahkan penelitian membuktikan banyak sekali negara yang melakukan plagiasi sehingga menimbulkan pencabutannya artikel yang negara tersebut terbitkan dan juga penyumbang terbanyak jurnal plagiasi itu sendiri dari biomedis, lebih parahnya hanya sedikit sekali artikel maupun jurnal yang diketahui melakukan plagiasi dicabut [2]. Banyak cara atau metode yang efektif untuk mencocokkan antar *query*, mulai dengan metode *Jaccard*, *Dice*, *Cosine Similarity* dll untuk menghitung atau mencari sebuah kemiripan, *Cosine* dan *Dice Similarity* biasa digunakan pada kasus *Web Crawling*, *Plagiarism*, *Search Engine*, mengukur/menghitung kemiripan pada sebuah dokumen serta program yang memiliki kaitan dengan *Information Retrieval System* [3] [4].

Disini peneliti menggunakan *Cosine* dan *Dice Similarity* dikarenakan hasil penelitian menunjukkan bahwa metode *Dice Similarity* dengan menggunakan *Query Expansion* memiliki nilai *recall* paling tinggi yaitu sebesar 85,357% dibanding metode lainnya. Metode *Cosine Similarity* menggunakan *Query Expansion* memiliki nilai *precision* yang tinggi dibanding dengan metode lainnya, yaitu sebesar 10,041% dan nilai *f-measure* yang tinggi dibandingkan metode lainnya yaitu sebesar 17,061%.[3]

Pada penjelasan diatas dijelaskan bahwa betapa pentingnya sebuah sistem pendeteksi kemiripan dan dengan ini peneliti melakukan perbandingan algoritma pada *Cosine* dan *Dice Similarity* dalam menghitung kemiripan dokumen. *Dice Similarity* adalah metode tentang mengukur tingkat kedekatan maupun kesamaan (*Similarity*) *term* dengan cara membobotkan *term*. Dokumen dipandang sebagai sebuah vektor yang memiliki *magnitude* (jarak) dan *direction* (arah) [5]. Pada tahun 2016 Fatkhul Amin, Purwatiningtyas, Edy Winarno melakukan sebuah penelitian berjudul Rancang Bangun Sistem Temu Kembali Informasi (*Information Retrieval System*) Dokumen Berbahasa Jawa menggunakan Metode *Dice Similarity* untuk tujuan mempermudah melakukan pencarian dokumen teks berbahasa Jawa penelitian itu menghasilkan rata-rata *recall* = 0,04 dan rata-rata *precision* = 0,83 [5].

Selain metode *Dice Similarity* ada juga metode lain yang menghitung kemiripan antara *query* dengan dokumen seperti *Cosine Similarity*. Terdapat penelitian yang telah dilakukan mengenai *Cosine Similarity*, yaitu penelitian yang dilakukan oleh Viko Basmalah Wicaksono, dkk,[6] yang memberi nilai rata-rata *precision* 44,82983% dan *recall* 99,08165% dalam uraian latar belakang ini peneliti melakukan penelitian berjudul **“Perbandingan Algoritma *Cosine* Dan *Dice Similarity* Dalam Menghitung Kemiripan Dokumen”**. Pada penelitian ini peneliti melakukan perbandingan algoritma metode *Cosine Similarity* dan *Dice Similarity* dalam menghitung kesamaan di dalam dokumen dan dengan menggunakan sistem perbandingan similaritas menggunakan *Cosine Similarity* dan *Dice Similarity* untuk menentukan bahwasanya dokumen yang diperiksa terdapat kemiripan atau tidak, secara mudah dan efisien.

1.2 Perumusan Masalah

1.2.1 Masalah Penelitian

Berdasarkan latar belakang diatas maka permasalahan yang dilakukan pada penelitian ini adalah membuat sistem perbandingan dalam mengukur similaritas metode tersebut.

1.2.2 Metode Usulan

Metode yang diusulkan yaitu penerapan Algoritma *Cosine similarity* dan *Dice Similarity* untuk menghitung kemiripan dokumen.

1.2.3 Pertanyaan Penelitian

Bagaimana hasil perbandingan similaritas dan waktu untuk menghitung

kemiripan dokumen menggunakan *Cosine Similarity* dan *Dice Similarity*.

1.3 Tujuan dan Manfaat Penelitian

1.3.1 Tujuan Penelitian

Dari pertanyaan penelitian di atas maka dengan itu tujuan dari penelitian ini adalah mengetahui perbandingan similaritas dan waktu pencarian kemiripan menggunakan *Cosine* dan *Dice Similarity* pada dokumen yang diuji dan juga membuat sistem yang membandingkan 2 metode tersebut, sehingga bisa digunakan di pada penelitian lanjutan.

1.3.2 Manfaat Penelitian

1. Adapun manfaat penelitian ini adalah mengetahui mana metode yang lebih baik antara *cosine* dan *dice* dalam mencari kemiripan dokumen.
2. Aplikasi ini berguna untuk untuk semua orang baik itu komunitas maupun badan akademisi dan juga bisa berguna untuk pengembangan penelitian maupun penelitian selanjutnya.

1.4 Batasan Masalah

Adapun batasan masalah pada penelitian ini adalah:

1. Dokumen abstrak yang digunakan diambil dari situs *PTA Universitas Trunojoyo Madura* 100 dokumen pada tanggal 1 Oktober 2020
2. Pencarian hanya bisa digunakan/dilakukan untuk pencarian teks berbahasa Indonesia.
3. Sistem dibuat berbasis web.
4. Kemiripan yang dicari hanya abstrak pada dokumen yang diinginkan.

1.5 Metodologi Penelitian

Berikut tahapan-tahapan penelitian yang dilakukan prosesnya adalah sebagai berikut:

a. Studi Literatur

Pada bagian ini peneliti mencari informasi yang berkaitan dengan pembahasan, perkembangan penelitian maupun perkembangan metode yang digunakan pada penelitian ini. salah satunya dengan cara mengumpulkan literatur berupa jurnal dan paper, khususnya yang mengangkat *Nazief* dan *Adriani*, *Cosine* dan *Dice Similarity* serta apapun yang membahas tentang konsep *Similarity*.

b. Pengumpulan Data

Data yang diambil oleh peneliti berupa data dokumen yang berada pada situs Portal Artikel Tugas Akhir Kampus Universitas Trunojoyo Madura (<https://pta.trunojoyo.ac.id/>).

c. Perancangan Sistem

Pada bagian ini peneliti menjelaskan bagaimana suatu sistem dibuat agar mempermudah peneliti untuk melakukan implementasi sistem, sehingga peneliti dapat mengevaluasi serta menganalisa hasil akhir yang didapatkan pada studi literatur yang diteliti.

d. Implementasi Sistem

Implementasi sistem adalah tahap dimana pembuat dan pengembangan dirancang sesuai dengan rancangan pada tahap sebelumnya. Sistem ini dibangun menggunakan basis *web* berbahasa pemrograman *PHP* dengan *user interface* untuk pengguna.

e. Pengujian Sistem

Pengujian dilakukan terhadap sejumlah input pengguna, mulai uji coba ke-1 sampai ke-*n*. Dari masing-masing pengujian dapat diperoleh nilai selisih kemiripan dari metode *Cosine* dan *Dice Similarity* terhadap nilai kemiripan dari pihak ke 3 yaitu aplikasi *Plagiarism Checker X* sebagai pembanding.

f. Dokumentasi dan Penyusunan Laporan

Dokumentasi dan penyusunan pada laporan ini yang dilakukan selama penelitian ini berlangsung. Sehingga penyusunan laporan pada penelitian ini dapat dilakukan sesuai dengan sistematika penulisan penelitian.