ISSN 2460-1160

IMPLEMENTASI METODE DICE SIMILARITY DALAM PERANCANGAN SISTEM REKOMENDASI ARTIKEL BERITA

Rudy Ariyanto [1], Yoppy Yunhasnawa [2], Muhammad Iqbal Rofikurrahman [3]

[1,2,3] Program Studi Teknik Informatika, Jurusan Teknologi Informasi, Politeknik Negeri Malang
[1.1]ariyantorudy@gmail.com, [2]yunhasnawa@polinema.ac.id, [3]admin@iqbalcakep.com

***Abstrak*— Brilio.net merupakan sebuah perusahaan media penyedia artikel yang sudah cukup dikenal di Indonesia. website ini menyediakan artikel tentang gaya hidup, berita viral, teknologi dan artikel-artikel jenaka yang sering diperbincangkan oleh banyak netizen di Indonesia, dan memiliki jumlah pengunjung yang banyak dengan rata-rata sepuluh ribu pengunjung setiap harinya, brilio.net belum menerapkan sistem rekomendasi (*content based filtering*) yang berdasarkan sesuai dengan tingkat kemiripan artikel yang sedang diakses oleh pengguna, hal ini dapat mengurangi ketertarikan pengguna untuk membaca artikel lainnya, karena artikel rekomendasi yang ada hanya berdasarkan artikel dengan jumlah akses terbanyak saja sehingga tidak ada peningkatan dalam statistik website di bagian lama pengguna mengakses. Brilio.net membutuhkan sebuah sistem rekomendasi artikel berita, dengan proses perhitungan yang lebih cepat tanpa terpengaruh dengan banyaknya pengguna yang mengakses dan dapat menampilkan artikel yang mirip sesuai dengan artikel yang diakses oleh pengguna. Metode *dice similarity* merupakan metode yang digunakan untuk mencari tingkat kemiripan pada sebuah dokumen dan tidak membutuhkan komputasi yang terlalu berat. Sehingga dengan adanya sistem rekomendasi artikel berita menggunakan metode *dice similarity* maka situs brilio.net dapat memberikan rekomendasi artikel berita yang mirip dengan artikel lain dan cepat tanpa terpengaruh dengan jumlah pengguna yang mengakses, dan hasil dari penelitian kali ini metode *dice similarity* kurang cocok dalam sistem rekomendasi artikel berita dengan nilai proses waktu 151 detik untuk 100 artikel dengan kinerja sistem memperoleh rata-rata precision 53,5% dan recall 46,5%.**

***Kata kunci: Content based filtering, Dice similarity,* Sistem rekomendasi.**

I. PENDAHULUAN

Artikel merupakan karangan tertulis yang panjangnya tidak dapat ditentukan, dimana tujuannya untuk menyampaikan gagasan maupun fakta dengan maksud meyakinkan, mendidik, ataupun menghibur [1]. Di era modern, artikel sering ditulis di berbagai media online seperti website. Artikel berita dapat dibagi menjadi berbagai kategori sesuai pembahasannya. Dalam sebuah kategori pastinya akan memiliki jumlah artikel yang sangat banyak. Sehingga pengkategorian artikel tidak dapat digunakan sebagai acuan sarana untuk mempermudah pengguna dalam memilih artikel yang tepat. Sistem rekomendasi hadir sebagai sarana agar pengguna dapat lebih mudah memilih artikel yang mirip dengan artikel lain sesuai yang dengan pengguna akses, tanpa perlu harus melakukan pencarian di kolom pencarian website.

Sistem rekomendasi merupakan sistem yang sering diterapkan oleh website penyedia artikel, adanya sistem rekomendasi dapat memberikan dampak yang baik terhadap website. Dalam sistem rekomendasi tingkat kemiripan terhadap artikel yang direkomendasikan harus diperhatikan, agar pengguna dapat memberikan respon yang baik pada statistik website dan pada artikel yang telah diakses. Sistem rekomendasi terdapat berbagai teknik dalam penerapannya, yaitu terdiri dari *content based filtering*, *collaborative filtering*, dan *hybrid filtering*.

Brilio.net merupakan sebuah perusahaan media penyedia artikel yang sudah cukup dikenal di Indonesia. website ini menyediakan artikel tentang gaya hidup, berita viral, teknologi dan artikel-artikel jenaka yang sering diperbincangkan oleh banyak netizen di Indonesia, dan memiliki jumlah pengunjung yang banyak dengan rata-rata sepuluh ribu pengunjung setiap harinya, brilio.net belum menerapkan sistem rekomendasi (*content based filtering*) yang berdasarkan sesuai dengan tingkat kemiripan artikel yang sedang diakses oleh pengguna, hal ini dapat mengurangi ketertarikan pengguna untuk membaca artikel lainnya, karena artikel rekomendasi yang ada hanya berdasarkan artikel dengan jumlah akses terbanyak saja, sehingga tidak ada peningkatan dalam statistik website dibagian lama pengguna mengakses.

Pada praktik sebelumnya banyak sistem rekomendasi artikel yang telah digunakan, namun sistem rekomendasi yang ada beroperasi kurang cepat, dikarenakan alur sistem yang akan melakukan contoh gaya penulisan/format tersedia. Penulis perlu membuat *style* tersendiri untuk menyamakan format komponen tersebut. operasi pengecekan kemiripan artikel disaat setiap pengguna mengakses, sehingga semakin banyak pengguna dan artikel yang diakses, semakin lama juga proses perhitungannya.

Brilio.net membutuhkan sebuah sistem rekomendasi artikel berita, dengan proses perhitungan yang lebih cepat tanpa terpengaruh dengan banyaknya pengguna yang mengakses dan dapat menampilkan artikel yang mirip sesuai dengan artikel yang diakses oleh pengguna. Sebuah artikel perbandingan menyebutkan bahwa metode *dice similarity* merupakan metode yang digunakan untuk mencari tingkat kemiripan pada sebuah dokumen dan tidak membutuhkan komputasi yang terlalu berat [2]. Sehingga dengan adanya sistem rekomendasi artikel berita menggunakan metode dice atau jaccard maka situs brilio.net dapat memberikan rekomendasi artikel berita yang mirip dengan artikel lain dan

cepat tanpa terpengaruh dengan jumlah pengguna yang mengakses.

II. LANDASAN TEORI

2.1 Content Based Filtering

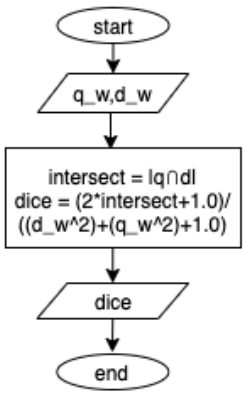
Sistem rekomendasi berbasis konten (Content-based Recommendation System) menggunakan ketersediaan konten (sering juga disebut dengan fitur, atribut atau karakteristik) sebuah item sebagai basis dalam pemberian rekomendasi [3]. Sistem ini cocok digunakan untuk membuat rekomendasi yang tidak membutuhkan histori atau aktifitas pengguna sebelumnya.

Perbedaan metode Content Based Filtering dengan collaborative filtering adalah persyaratan yang harus dipenuhi yaitu rekaman hasil pengguna selama mengakses sehingga bisa dikelompokkan tema atau judul artikel apa saja yang pengguna sukai. Oleh sebab itu metode collaborative filtering ini tidak dapat diterapkan pada studi kasus ini.

baku dan memiliki arti sendiri, tanpa adanya imbuhan atau kata bantu.

2.2 Dice Similarity

Dice Similarity adalah metode untuk melihat tingkat kedekatan atau kesamaan (similarity) term dengan cara pembobotan term. Dokumen dipandang sebagai sebuah vektor yang memiliki *magnitude* (jarak) dan direction (arah). Kelebihan dari metode ini adalah proses komputasi yang tidak terlalu berat namun akurat dan dapat menentukan prioritas antara precision dan recall berdasarkan kebutuhan melalui nilai pada persamaan (2).



Gambar 6. Flowchart Tokenisasi

Untuk perhitungan metode dice similarity ini dapat dilakukan dengan rumus sebagai berikut :

$$dice = \frac{2 * |d1 \cap d2| + 1.0}{(d2^2) + (d1^2) + 1.0} \quad (2).$$

Dimana d1 merupakan dokumen pertama atau dokumen yang sedang diakses oleh pengguna, d2 adalah dokumen pembanding, pada sistem ini dokumen pembanding adalah semua artikel yang telah tercrawling oleh sistem.

III. METODE PENELITIAN

3.1 Data

Data yang akan diolah merupakan data yang berasal dari artikel yang telah dibuat sebelumnya pada situs brilio.net (data uji), pengolahan data berfokus kepada isi artikel seperti judul, gambar, isi konten, tanggal terbit dan alamat URL. Dalam proses perhitungan algoritma, semua data kecuali alamat URL, gambar, dan tanggal terbit akan digabung menjadi satu dokumen untuk dibandingkan dengan dokumen yang telah dibuat sebelumnya.

3.2 Pengambilan Data

Metode yang digunakan dalam proses pengambilan data adalah metode crawling, dimana metode ini juga dikenal dengan metode scrapping yang intinya melakukan ekstraksi sebuah halaman pada situs tertentu untuk memperoleh data yang dibutuhkan, khususnya pada penelitian kali ini adalah situs brilio.net.

Metode ini memiliki tingkat kesulitan tersendiri tergantung situs yang akan dilakukan ekstraksi. Semua data yang diperoleh adalah data legal, dan semua data yang tampil saat membuka situs secara manual adalah data yang dapat dilakukan proses crawling.

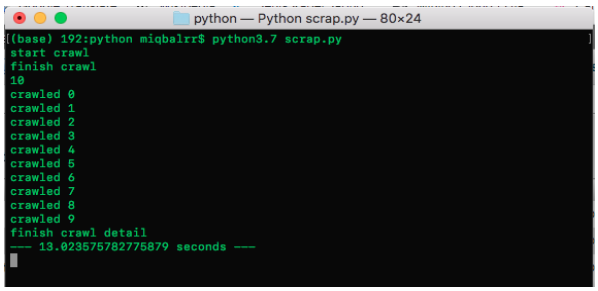
IV. PENGUJIAN

4.1 Pengujian Fungsional

Sistem rekomendasi artikel memiliki banyak fungsi yang dijalankan, sehingga pengujian dilakukan agar alur sistem secara keseluruhan dapat berjalan dengan baik.

a. Pengujian Crawling

Fungsi *crawling* berguna untuk mengambil data artikel yang terdapat pada situs brilio.net, pengujian *crawling* dilakukan dengan cara mengambil data sebanyak sepuluh artikel.

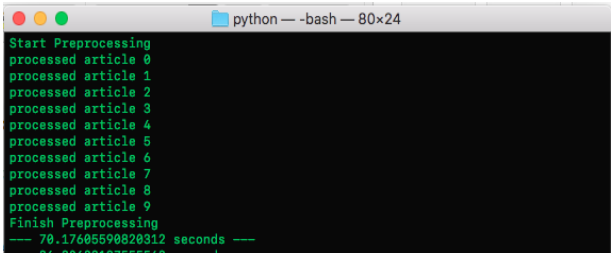


Gambar 15. Pengujian Crawling

Pengujian sukses dengan melakukan pengambilan sepuluh data artikel pertama pada situs brilio.net. waktu yang dibutuhkan untuk melakukan *crawling* dengan *default network caching* sebanyak 13 – 15 detik bergantung kepada koneksi internet yang digunakan.

b. Pengujian *Preprocessing*

Data yang diuji pada fungsi *preprocessing* merupakan sepuluh data yang telah ter*crawling* sebelumnya dan telah tersimpan kedalam sebuah variabel.

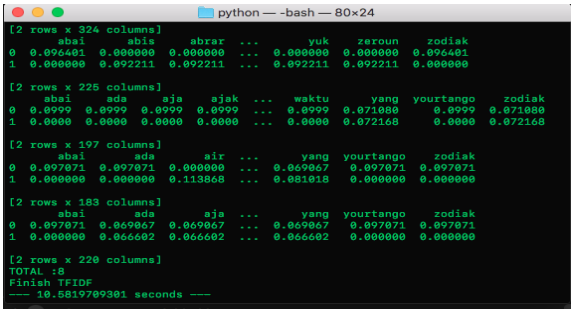


Gambar 16. Pengujian *Preprocessing*

Rentan waktu yang diberikan untuk preprocessing sepuluh data adalah sekitar 60 – 70 detik. hasil preprocessing akan tersimpan dalam database dan tidak akan dilakukan preprocessing ulang untuk data yang telah tersimpan.

c. Pengujian Pembobotan TF IDF

Pengujian pembobotan TF IDF dilakukan dengan menguji hasil pembobotan dari artikel yang telah dilakukan *preprocessing*. Proses pembobotan dapat dilakukan pada persamaan (1) dan tabel 1



Gambar 17. Pengujian Pembobotan

Total artikel yang dilakukan pembobotan ada 8 artikel yang dibandingkan masing-masing dua artikel. Proses pembobotan menghabiskan rentan waktu antara 10 – 15 detik, bergantung pada jumlah artikel yang diproses.

d. Pengujian Perhitungan

Pengujian perhitungan dilakukan setiap metode yang diterapkan dalam sistem, berikut adalah pengujian metode dice.

Data sampel dalam proses pengujian ini :

Dokumen1 / d1 = “Indonesia neraga maju” sebagai *query*
Dokumen2 / d2 = “Indonesia negara berkembang”

TABEL 4.1 PEMBOBOTAN TFIDF

	tf				W=tf*idf	
	d1	d2	df	idf	d1	d2
indonesia	1	1	2	0,501	0,501	0,501
negara	1	1	2	0,501	0,501	0,501
maju	1	0	1	0,704	0,704	0
berkembang	0	1	1	0,704	0	0,704
				Sum/total	1,706	1.706

Pada tabel 1 dapat diperoleh bobot untuk dokumen 1 dan dokumen 2 adalah sama sebesar 1,706, setelah proses pembobotan selesai dilanjut dengan proses persamaan (2).

$$intersection = (0{,}501 + 0{,}501) * (0{,}501 + 0{,}501) = 1{,}004$$

$$dice = \frac{2.0 * intersection + 1.0}{((1{,}706^2) + (1{,}706^2) + 1.0)}$$

$$dice = \frac{3{,}008}{6{,}8208} = 0{,}4405$$

Berikut adalah hasil pengujian dari sistem menggunakan data sampel serupa, menunjukkan nilai kemiripan sebesar 44%.

Gambar 18. Hasil Dice Dari Sistem

V. HASIL DAN PEMBAHASAN

Penerapan sistem rekomendasi artikel sangat membutuhkan sebuah proses yang cepat dan memiliki kinerja sistem yang baik, sehingga penelitian kali ini menerapkan dua metode berbeda. Keakuratan dapat dinilai dari kecocokan antara dua metode.

5.1 Hasil precision dan recall

Setelah melakukan pengujian dengan 2 kata kunci berdasarkan data pada tabel 5.2 dan tabel 5.3 hasil yang diperoleh pada kata kunci pertama “meme lucu” adalah sebagai berikut:

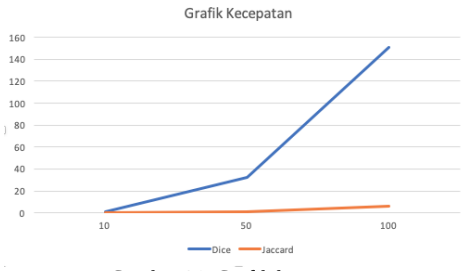
Metode dice similarityprecision=3/(3+3)*100=50%
recall= 3/(3+2)*100=60%.

Perbedaan yang sangat signifikan ditunjukkan pada bagian kecepatan, perbedaan kecepatan antara dua metode *dice* dan *jaccard* terpaut sangat jauh, dari 100 artikel yang diproses metode *jaccard* bahkan melakukan perhitungan kurang dari 10 detik. hal ini sangat wajar karena metode jaccard pada sistem ini tidak melalui proses pembobotan.
Proses pembobotan menghabiskan waktu sekitar 42 detik sehingga jika diakumulasikan metode dice tanpa pembobotan dengan persamaan berikut:

total waktu−waktu pembobotan = waktu tanpa pembobotan (6).

151 – 42 = 109

Akumulasi proses metode *dice* tanpa pembobotan sekitar 109 detik masih lebih jauh lebih cepat dibandingkan dengan metode *jaccard* tanpa pembobotan.



Gambar 21. Grafik kecepatan

Berdasarkan gambar 21 dapat diperoleh data bahwa metode *jaccard* sangat jauh lebih cepat dibandingkan dengan metode *dice*. Dari 10 artikel metode *dice* membutuhkan waktu 1,4 detik

Berikut adalah beberapa faktor yang mempengaruhi lambat nya metode *dice similarity*:

1. Metode *dice* membutuhkan proses pembobotan TF.IDF
2. Metode *dice* melakukan lebih banyak perhitungan
3. Semakin banyak dokumen yang dibandingkan proses akan semakin lama dikarenakan pada sistem ini proses perbandingkan dilakukan pada setiap dokumen satu persatu.

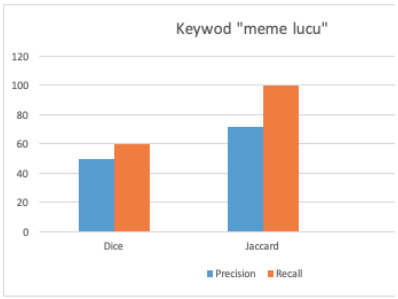
Pembahasan kecepatan juga hanya berfokus kepada proses kecepatan metode, tidak berpengaruh terhadap kecepatan proses *crawling* hal ini dikarenakan kecepatan proses *crawling* dapat dipengaruhi dari beberapa faktor luar seperti kecepatan koneksi internet yang digunakan, serta spesifikasi komputer yang sedang digunakan sangat berpengaruh terhadap hasil kecepatan.

Pembahasan lain adalah bagian kinerja sistem, hasil perbandingan kesamaan disetiap index sangatlah beragam, namun kesimpulan yang dapat diambil adalah dari 129 artikel dengan maksimal index ke-5 jumlah hasil rekomendasi artikel mirip masih menunjukkan persentase diatas 50%, yang artinya kedua metode memberikan rekomendasi yang cukup mirip jika dilihat berdasarkan id artikel yang direkomendasikan. Meskipun demikian rata-rata persentase tingkat kemiripan artikel pada metode jaccard dan dice berbeda.

Gambar 22. persentase kemiripan

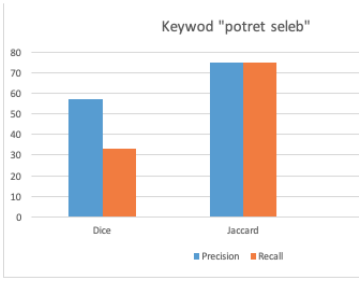
Berdasarkan gambar 22 metode dice dan jaccard memberikan hasil kemiripan artikel yang sama, namun memberikan nilai persentase berbeda yang nilai nya lebih baik diperoleh dari metode jaccard.

Dengan metode pengukuran kinerja sistem menggunakan *precision* dan *recall* dapat memberikan kesimpulan yang lebih meyakinan, metode *dice* memberikan hasil pada kata kunci pertama sebesar 50% dan 60% untuk kata kunci kedua sebesar 57.1% dan 33.3%.



Gambar 23 grafik hasil keyword pertama

Berdasarkan gambar 23 hasil *recall* dari masing-masing metode memberikan nilai persentase yang cukup tinggi, dan *precision* yang lebih rendah meskipun tidak dibawah 50%.



Gambar 24 grafik hasil keyword kedua

Pada gambar 24 metode *dice* menunjukkan penurunan drastis pada hasil *recall* menggunakan kata kunci kedua

VI. KESIMPULAN DAN SARAN

Dari penelitian yang telah dilakukan dapat diambil kesimpulan sebagai berikut:

Hasil pengujian dari segi kecepatan metode jaccard jauh lebih baik dengan hanya membutuhkan waktu 5,6 detik untuk 100 artikel dan berdasarkan hasil pengujian kinerja sistem menggunakan metode precision dan recall, metode dice memperoleh nilai dengan rata-rata persentase precision 53,5% dan recall 46,5%, Sehingga metode dice adalah metode yang kurang cocok digunakan untuk implementasi sistem rekomendasi pada artikel berita berdasarkan waktu proses dan kinerja sistemnya.

Dengan selesainya Implementasi Metode Dice Similarity Dalam Perancangan Sistem Rekomendasi Artikel Berita, berikut adalah saran untuk pengembangan sistem lebih lanjut:

1. Proses crawling bisa dikembangkan untuk berbagai situs media. Sehingga artikel yang menjadi acuan untuk rekomendasi memiliki banyak sumber.

2. Melakukan optimasi pada bidang preprocessing, sehingga proses kemiripan bisa dilakukan berdasarkan arti / semantic

DAFTAR PUSTAKA

[1] Agus, M., Subali, P., Fatichah, C., & Informatika, D. (2019). Kombinasi Metode Rule-Based Dan N-Gram Stemming Untuk a Combination of Methods Rule-Based and N-Gram Stemming To Recognize Balinese Language Stemmer. *JTIIK : Jurnal Teknologi Informasi Dan Ilmu Komputer*, *6*(2). https://doi.org/10.25126/jtiik.201961105

[2] Dzulfikri, A. (2019). Perbandingan Metode Dice Similarity Dengan *Cosine Similarity* Menggunakan *Query* Expansion Pada Pencarian Ayatul Ahkam Dalam Terjemah Alquran Berbahasa Indonesia. Malang: UIN Maulana Malik Ibrahim Malang.

[3] Herwijayanti, B., Ratnawati, D. E., & Muflikhah, L. (2018). Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity. *Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(1), 306–312.