Building of Informatics, Technology and Science (BITS)
Volume 3, No 3, December 2021, Page 275−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i3.1058

Penerapan Metode N-Gram dan Cosine Similarity Dalam Pencarian Pada Repositori Artikel Jurnal Publikasi

Indra Gita Anugrah

Fakultas Teknik, Program Studi Sistem Informasi, Universitas Muhammadiyah Gresik, Gresik, Indoesia
Email: [1]indragitaanugrah@umg.ac.id
Email Penulis Korespondensi: indragitaanugrah@umg.ac.id
Submitted: **18/12/2021**; Accepted: **26/12/2021**; Published**: 31/12/2021**

Abstrak−Repositori digital merupakan salah satu sumber data dalam kebutuhan informasi manusia, khususnya pada sebuah organisasi. Dalam sebuah repositori digital tersimpan berbagai dokumen digital yang dapat dimanfaatkan oleh pengguna, contohnya adalah repositori jurnal publikasi. Setiap harinya artikel publikasi dalam repositori mengalami perkembangan besarnya ratusan bahkan ribuan jumlahnya, selain itu jurnal publikasi biasanya terdiri berbagai format dan bahasa Hal ini akan menyebabkan hasil pencarian memiliki tingkat relevansi yang relatif rendah. Untuk mengoptimalkan hasil pencarian dewasa ini penerapan information retrieval system pada sebuah repositori menjadi penting. Praproses merupakan salah satu tahapan yang terpenting dari pengembangan retrieval system, terutama dalam proses pemilihan algoritma stemming untuk menghasilkan kata dasar (term) yang nantinya digunakan dalam menentukan tingkat kemiripan antara query dan dokumen pada sebuah proses pencarian. N-Gram merupakan metode dekompososi karakter dari sebuah string yang dapat digunakan untuk menganalisa kata atau kalimat. Hasil analisa dapat diketahui kata atau kalimat dari bahasa inggris atau bahasa indonesia, sehingga algoritma stemming dipilih berdasarkan jenis bahasa yang telah diketahui. Cosine Similarity merupakan metode untuk menghitung besar tingkat kemiripan, dimana akan dihitung besaran sudut yang merepresentasikan vector query dan vector dokumen. Pada penelitian ini akan dibangun repositori yang menerapkan retrieval systems menggunakan N-Gram dan Cosine Similarity, kemudian akan dihitung kinerja system dimana didapatkan total rata-rata akurasi untuk untuk query berbahasa Indonesia dan query berbahasa inggris sebesar 0,967, presisi sebesar 0,851 sedangkan rata recall didapatkan hasil 0,869.

Kata Kunci: NGram; Cosine Similarity; Repositori Digital; Temu Kembali Informasi; Similarity

Abstract−Digital repository is one source of data in human information needs, especially in an organization. In a digital repository, various digital documents are stored that can be used by users, for example, a publication journal repository. Every day the published articles in the repository grow in the hundreds or even thousands in number, besides publication journals usually consist of various formats and languages. This will cause the search of results relatively low level of relevance. To optimize search results today, the application of an information retrieval system in a repository is important. Preprocessing is one of the most important stages of the development of a retrieval system, especially in the process of selecting a stemming algorithm to generate basic words (terms) which will later be used in determining the level of similarity between queries and documents in a search process. N-Gram is a method of character decomposition from a string that can be used to analyze words or sentences which are words or sentences from what language will later affect the determination of the stemming algorithm. Cosine Similarity is a method to calculate the level of similarity, which will calculate the angle that represents the query vector and the document vector. In this study, a repository will be built that implements retrieval systems using N-Gram and Cosine Similarity, then the system performance will be calculated where the average total accuracy for Indonesian-language queries and English-language queries is 0.967, precision is 0.851 while the average recall is obtained. 0.869.

Keywords: NGram; Cosine Similarity; Digital Repository; Retrieval Information System; Similarity

1. PENDAHULUAN

Perkembangan Teknologi Informasi dan Komunikasi memberikan dampak yang besar dalam kehidupan manusia, salah satunya adalah penggunaan koleksi dokumen secara digital. Repositori merupakan contoh koleksi dokumen digital yang berisi berbagai macam koleksi, salah satunya adalah kumpulan artikel penelitian dan jurnal. Artikel penelitian dan jurnal yang tersimpan dalam repositori didapatkan dari berbagai sumber baik dari hasil artikel penelitian dari mahasiswa maupun dosen atau dari peneliti dari pihak luar, baik artikel jurnal menggunakan berbahasa Indonesia maupun berbahasa Inggris. Penggunaan repositori digital membantu pengguna dalam menemukan literasi yang diinginkan baik mahasiswa maupun dosen. Pemanfaatan repositori digital seringkali juga digunakan untuk menemukan referensi dalam penulisan artikel ilmiah. Dalam pengembangan repositori yang digunakan untuk menyimpan kumpulan artikel publikasi ilmiah. Sebuah repositori biasanya tersimpan ratusan hingga ribuan dokumen dengan bermacam-macam format dan bahasa [1], hal ini mengakibatkan permasalahan pada proses pencarian sehingga menghasilkan tingkat relevansi yang rendah.

Pada umumnya pencarian di repositori yang menggunakan *information retrieval*, pada proses pencarian harus melalui beberapa tahapan, salah satunya adalah tahap *stemming*. *Stemming* merupakan proses mengubah kalimat menjadi kata dasar dengan menghilangkan imbuhan [2]. Pada repositori banyak artikel ilmiah ditemukan dalam bahasa Indonesia tetapi abstraknya berbahasa bahasa Inggris, sehingga diperlukan pendeteksian bahasa untuk menentukan algoritma stemming yang akan digunakan dikarenakan algoritma stemming bahasa Indonesia dan bahasa Inggris memiliki perbedaan dalam algoritma sesuai tata bahasa. Pendekatan yang dapat digunakan dalam pendeteksian bahasa salah satunya adalah menggunakan N-Gram. N-Gram merupakan pemecahan dan kombinasi berdasarkan jumlah N karakter dari sebuah kata atau string yang kemudian dibandingkan dengan kata yang telah disimpan pada koleksi kata

Building of Informatics, Technology and Science (BITS)
Volume 3, No 3, December 2021, Page 275−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i3.1058

(corpus) [3]. Selain pendeteksian bahasa, untuk mengoptimalkan pencarian diperlukan sebuah mekanisme mengukur kemiripan anatara query yang diinputkan dengan koleksi dokumen dalam repositori.

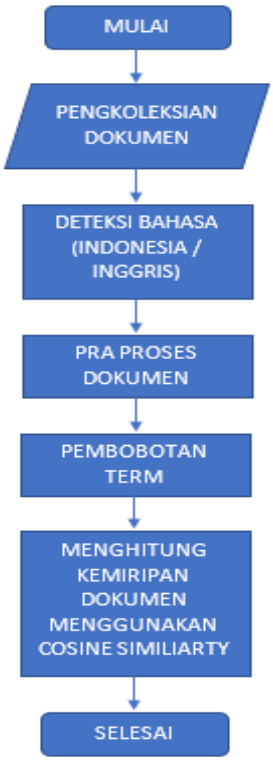
Metode Cosine Similarity merupakan salah satu metode yang digunakan untuk mengukur kemiripan dokumen dengan membandingkan antara vector query dengan vector dokumen, dimana sebelumnya query dan dokumen direpresentasikan dalam model ruang vector (Vektor Space Model). Sebagaimana penelitian [4] dengan menggunakan cosine similarity dan pembobotan TF-IDF diperoleh tingkat akurasi sebesar 98%. Selanjutnya [5] mencoba membandingkan Jaccard dan cosinus similarity pada uji kemiripan dokumen, dengan hasil yang menunjukkan bahwa uji similaritas, akurasi menggunakan cosine similarity sebesar 0.949808 unggul dibanding Jaccard dengan 0.949077l. Untuk deteksi bahasa dalam [6] penelitian tentang sistem deteksi bahasa menggunakan N-Gram, kinerjanya cukup baik untuk mendeteksi Bahasa dengan rata-rata F-measure 0,93.

Pada penelitian ini akan dilakukan penerapan algoritma N-Gram yang digunakan untuk mendeteksi bahasa dan metode Cosine Similarity dalam mengukur kemiripan antara query dan dokumen yang direpresentasikan dalam Vector Space Model (VSM) [7]. Kumpulan dokumen yang digunakan merupakan kumpulan artikel jurnal dan nantinya akan dicari kemiripan dengan query yang diinputkan pada proses pencarian suatu repositori sehingga dapat digunakan sebagai referensi dalam penulisan artikel jurnal. Tujuan dari implementasi N-Gram dan Cosine Similarity diharapkan dapat meningkatkan kinerja dari pencarian pada repositori digital.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Pada Gambar 1 merupakan alur penelitian yang akan digunakan, Adapun tahap yang akan dilakukan adalah sebagai berikut: pengkoleksian dokumen menggunakan scraping dokumen, deteksi bahasa menggunakan N-Gram, Praproses dokumen menggunkan algoritma stemming sastrawi untuk bahasa Indonesia sedangkan algoritma snowball untuk bahasa inggris, pembobotan term menggunakan TF-IDF yang akan di representasikan kedalam Vector Space Model (VSM), kemudian tahapan akhir adalah menghitung kemiripan dokumen menggunakan Cosine Similarity.



Gambar 1. Alur Penelitian

2.2 Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan hasil, Web Scraping atau Document Scraping adalah proses pengambilan dokumen HTML dari internet untuk mengambil data tertentu dari halaman untuk tujuan tertentu [8]. Dalam penelitian ini penulis menggunakan sumber dokumen artikel ilmiah dari internet pada website www.neliti.com yang menyediakan berbagai macam dokumen artikel ilmiah Gambar 2 merupakan alur proses Document Scraping, dimulai dari inputan kata kunci kemudian sistem akan mengolahnya untuk mendapatkan data dengan struktur yang diinginkan yaitu judul, abstrak, kata kunci, penerbit, kemudian data tersebut disimpan ke basis data sebagaimana penelitian yang dilakukan oleh [9].

Building of Informatics, Technology and Science (BITS)
Volume 3, No 3, December 2021, Page 275−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i3.1058

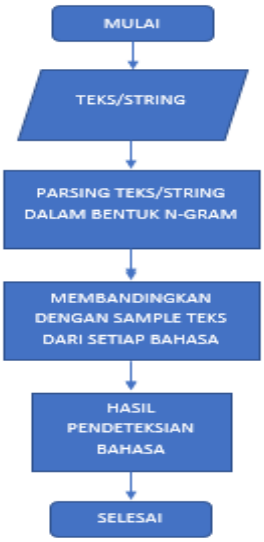
Gambar 2. Proses Pengambilan Data/ Document Scraping

2.3 Pendeteksian Bahasa

Penggunaan N-Gram untuk pendeteksian bahasa didasarkan pada asumsi bahwa pola distribusi N-gram suatu bahasa bersifat unik karena berkaitan dengan frekuensi penggunaan huruf, atau pasangan huruf, baik vokal maupun konsonan dari suatu bahasa yang umumnya berbeda dengan bahasa lain [6]. Gambar 3 merupakan alur dari pendeteksian Bahasa menggunakan N-Gram, N-Gram memiliki beberapa pendekatan dalam memotong karakter [10]. Untuk membantu dalam pengambilan potongan kata berupa karakter huruf tersebut, padding dilakukan dengan memberikan karakter "_" di awal dan akhir kata. Misalnya, kata "MAAF" dapat dipecah menjadi N-Gram berikut ("_" mewakili kosong) :

Tabel 1. Model N-Gram dan Pemecahan Karakter

N-Gram	Pemecahan Karakter
Uni-Gram	M, A, A, F
Bi-Gram	_M, MA, AA, AF, F_
Tri-Gram	_MA, MAA, AAF, AF_
Quad-Gram	_MAA, MAAF , AAF_
Quint-Gram	_MAAF, MAAF_



Building of Informatics, Technology and Science (BITS)
Volume 3, No 3, December 2021, Page 275−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i3.1058

Gambar 3. Proses Pendeteksian Bahasa Menggunakan N-Gram

N-Gram memiliki keunggulan dalam pengecekan sebuah string (string matching), salah satunya dikarenakan N-Gram memiliki sensitifitas yang rendah terhadap kesalahan penulisan [3]. Adapun karakterisitik dari N-Gram sebagai berikut:

1. Toleran terhadap kesalahan tekstual
2. Efisien karena memiliki algoritma yang sederhana dan waktu prosesnya cepat

Setelah teks diubah menjadi N-Gram, kemudian dicocokkan dengan data teks sampel pada masing-masing Bahasa dalam corpus dari setiap kamus bahasa. Dalam penelitian ini kami menggunakan dua buah corpus, yaitu corpus yang berisi kata-kata / term untuk bahasa Indonesia dan bahasa Inggris.

2.4 Dokumen Preprocessing

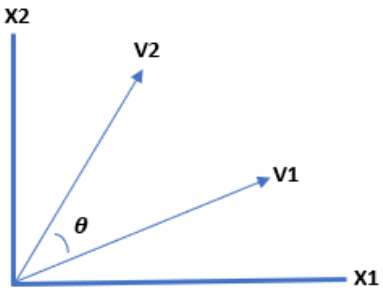
Tahap selanjutnya setelah bahasa terdeteksi adalah dokumen preprocessing. Preproses dokumen bertujuan membersihkan data dari noise, memiliki dimensi yang lebih kecil, dan lebih terstruktur sehingga dapat diolah lebih lanjut. Tahap preprocessing memiliki beberapa proses yaitu Case Folding, Stop Word Removal, Tokenizing, dan Stemming [2].

Adapun tahap dari Preprocessing Dokumen sebagai berikut :

1. Casefolding :
   Case folding merupakan proses pertama dalam serangkaian proses dokumen preprocessing. Proses ini bertujuan agar karakter huruf menjadi sama dengan cara huruf dalam dokumen diubah menjadi huruf kecil. Hanya huruf a sampai z yang diterima [11].
2. Stopword Removal :
   Stopword removal adalah tahap menghilangkan kata-kata yang tidak perlu dari teks. Stopwords merupakan kata yang tidak memiliki makna bila berdiri sendiri dan harus dibersihkan dalam pendekatan bag-of-words [12]. Pada tahap ini penulis menggunakan fungsi penghilangan stopword pada pustaka stemmer library.
3. Stemming.
   Pada tahap ini bertujuan untuk mencari kata dasar dengan membuang segala bentuk imbuhan. Stemming adalah proses pemetaan kata pada sebuah kalimat yang bermanfaat menjadi kata asli (tanpa awalan, akhiran, penyisipan, kombinasi) yang dieksekusi algoritma tertentu [13]. Tahap stemming dalam penelitian ini untuk bahasa Indonesia menggunakan algoritma stemmer sastrawi sebagaimana yang dilakukan oleh [14], sedangkan untuk bahasa Inggris menggunakan algoritma snowball stemmer.
4. Tokenisasi. Tokenisasi adalah proses membagi teks berupa kalimat atau paragraf dalam suatu dokumen menjadi token-token tertentu [15]. Pada tahap ini teks akan disusun berdasarkan syarat hasil stemming.

2.5 Vector Space Model (VSM)

Dokumen yang ada dalam repositori sistem temu kembali dokumen direpresentasikan sebagai vector. Sebuah dokumen terdiri dari beberapa kalimat dan setiap kalimat terdiri dari beberapa kata, sehingga dalam vector space model baik kalimat maupun query akan direpresentasikan sebagai vector.



Gambar 4. Representasi Jarak dan Vektor pada Cosine Similarity

2.6 Pembobotan Term

Pada tahap ini query pencarian dan dataset artikel ilmiah dilakukan pembobotan kata atau istilah untuk menghitung frekuensi kemunculan setiap kata (term frequency) query pencarian pada setiap artikel ilmiah dalam dataset. TF-IDF merupakan metode untuk memberikan bobot setiap term/ kata dasar dengan menghitung frekuensi kemunculan term pada setiap dokumen dalam pencarian informasi. Cara ini juga dikenal efisien, mudah dan memiliki hasil yang akurat [16]. Rumus pembobotan term weighting untuk penelitian ini menggunakan rumus TF-IDF adalah sebagai berikut.

$$w = tf\ x\ idf \quad (1)$$

Dimana :
tf : Jumlah term yang muncul pada setiap dokumen
idf : inverse dokumen frequency

Building of Informatics, Technology and Science (BITS)
Volume 3, No 3, December 2021, Page 275−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i3.1058

2.7 Cosine Similarity

Cosine Similarity merupakan salah satu metode pengukuran kesamaan dalam mekanisme sistem temu kembali dokumen. mengukur kesamaan antara dua dokumen atau teks [7]. Dalam proses pengukuran kesamaan Cosine Similarity menghitung nilai dari sudut yang dihasilkan [3]. Besaran atau nilai yang dihasilkan dari sudut vector antara 0 – 1, dimana semakin mendekati 1 maka query dan dokumen memiliki kemiripan yang besar dan semakin mendekati 0 maka memiliki kemiripan yang rendah. Cosine similarity dapat dihitung menggunakan Persamaan 2 sebagai berikut:

$$CosSim(d_i, q_i) = \frac{q_{i.} d_i}{|q_{i.} d_i|} = \frac{\sum_j^t = 1(q_{ij.}\, d_{ij})}{\sqrt{\sum_j^t 1(q_{ij.})^2 + \sum_j^t = 1(d_{ij.})^2}} \quad (2)$$

Keterangan :
q_{ij} : Term ke i untuk dokumen ke j dari q
d_{ij} : Term ke i untuk kueri ke j (keyword term)
t : Jumlah istilah j pada q atau d

2.8 Confusion Matrix

Pengukuran performa temu kembali sistem informasi dilakukan untuk melihat keberhasilan performansi pencarian oleh sistem [17]. Confusion Matrix dapat diajukan untuk pengukuran performansi, dimana representasi matriks dapat menggambarkan realitas dari data dan kolom dapat merepresentasikan prediksi dari sIstem atau sebaliknya. Confusion Matrix dapat direpresentasikan pada Tabel 2 berikut:

Tabel 1. Model N-Gram dan Pemecahan Karakter

	Predicted Negative	Predicted Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

Keterangan :
1. True Positive (TP) : Dimana data aktual positif dan hasil prediksi model juga positif.
2. True Negative (TN) : Dimana data aktual negatif dan hasil prediksi model juga negatif.
3. False Positive (FP) : Dimana data aktual negatif tetapi hasil prediksi model positif.
4. False Negatif (FN) : Dimana data aktual positif tetapi hasil prediksi model negatif.

Setelah ditentukan kelas dari confusion matrix maka performa model dapat diukur. Adapun pengukuran performa yang dapat dilakukan adalah :

1. Akurasi untuk mengetahui total keseluruhan data benar dengan menggunakan Persamaan 3

$$Akurasi = \frac{TP + TN}{Total\ Data} \quad (3)$$

2. Presisi untuk mengetahui kebenaran prediksi data yang benar dengan menggunakan Persamaan 4

$$Akurasi = \frac{TP}{FP + TP} \quad (4)$$

3. Recall untuk mengetahui seringnya sebuah model melakukan prediksi benar dengan menggunakan Persamaan 5

$$Recall = \frac{TP}{FN + TP} \quad (5)$$

3. HASIL DAN PEMBAHASAN

Pada bab ini ditampilkan hasil implementasi dari pencarian repositori artikel jurnal menggunakan N-Gram dan Cosine Similarity pada penelitian ini diawali dengan proses document scraping, deteksi bahasa menggunakan N-Gram, kemudian dilakukan praproses dokumen menggunkan algoritma stemming sastrawi untuk bahasa Indonesia

Building of Informatics, Technology and Science (BITS)
Volume 3, No 3, December 2021, Page 275−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i3.1058

sedangkan algoritma snowball untuk bahasa inggris, kemudian berikutnya dilakukan pembobotan term menggunakan TF-IDF dan akan di representasikan kedalam Vector Space Model (VSM), kemudian tahapan akhir adalah menghitung kemiripan dokumen menggunakan Cosine Similarity.

Data yang digunakan pada penelitian ini didapat dari proses scraping di website www.neliti.com sebanyak 100 data jurnal publikasi. Pada proses scraping akan diambil data isi teks dalam file, dengan memberikan batasan awal dan akhir yang telah ditentukan sesuai dengan struktur jurnal publikasi ilmiah, yaitu "Judul", "Abstrak", "Kata Kunci", dan "Penerbit". Berikut source code document scraping ditunjukkan pada Gambar 5, sedangkan hasil dari proses document scraping ditunjukan pada Gambar 8.

```
$source="artikel/$filename";
$output_folder="Penelitian/artikel"; if (!file_exists($foldOut))
{ mkdir($foldOut, 07 77, benar); } $a=passthru("pdftohtml $source
$foldOut/$filename",$b);
$myfile =fopen("Repositori/artikel/$filename$s.html" ,"r") or error($filename); $teks=
strtolower(fread($filesaya,ukuranfile("Penelitian/artikel/$filename$s.html")));
```

Gambar 5. Script Document Scraping

Gambar 6. Pencarian Data Artikel Jurnal Pada website www.neliti.com

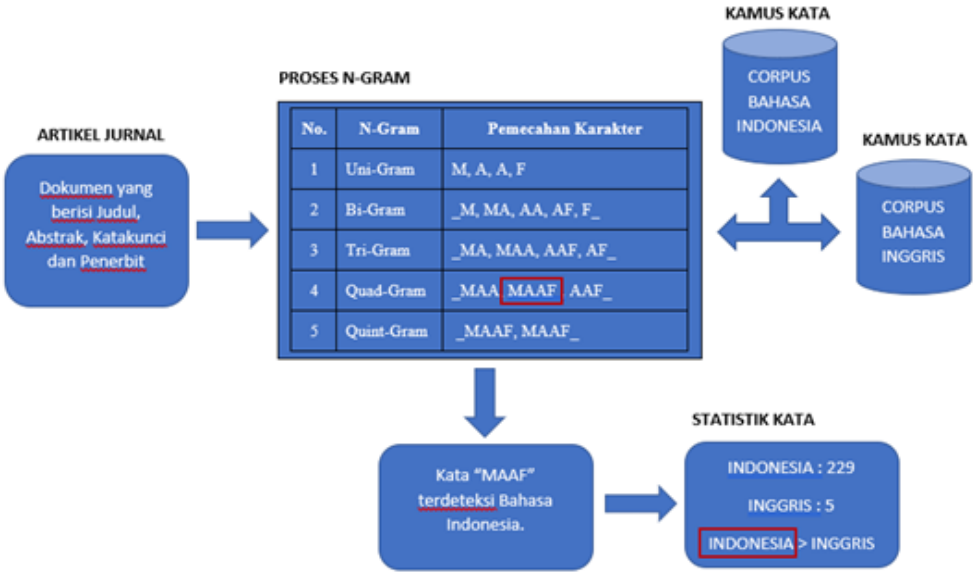
Pada Gambar 6 merupakan Proses Document Scraping pada aplikasi dimana, terlebih dahulu memasukan keyword dari website www.neliti.com. Setelah dilakukan pencarian maka akan menghasilkan informasi data dari artikel jurnal yang siap dilakukan proses scraping seperti yang ditunjukan pada Gambar 7.

Gambar 7. Proses Document Scraping

Struktur	Teks Hasil Document Scraping
Judul	Kajian Efektivitas Sistem Informasi Akademik Universitas Mercu Buana
Abstrak	Penerapan suatu sistem informasi pada organisasi dapat dikatakan berhasil apabila sistem informasi tersebut sukses dalam pelaksanaannya. Salah satu indikator kesuksesan sistem informasi adalah jika sistem informasi tersebut efektif penggunaannya. Model Kesuksesan Sistem Informasi DeLone dan McLean. secara teori dan praktek telah banyak didukung oleh beberapa peneliti untuk mengukur kesuksesan sistem informasi pada organisasi. Penelitian ini bertujuan untuk melakukan kajian kesuksesan sistem informasi akademik pada Universitas Mercu Buana dengan menggunakan model kesuksesan sistem informasi DeLone dan Mclean data dikumpulkan melalui survei kepada mahasiswa/wi, dosen dan karyawan pengguna sistem informasi akademik. Pengambilan sampel dalam penelitian ini menggunakan kuesioner yang diisi oleh 120 responden. Metode pengolahan data dianalisis dengan Pemodelan Persamaan Struktural (Structured Equation Modeling) dan memanfaatkan perangkat lunak AMOS 7.0. Hasil penelitian didapatkan bahwa secara umum kualitas informasi (KI) dan penggunaan (P) tidak mempengaruhi efektivitas pemanfaatan sistem informasi akademik pada kampus Universitas Mercu Buana, sedangkan faktor â€“ faktor yang mempengaruhi adalah kualitas sistem (KS), kualitas pelayanan (KP) dan kepuasan pengguna (KPG). Dan hasil yang didapat berdasarkan kelompok pengguna Dosen/Karyawan adalah kualitas sistem (KS), kualitas pelayanan (KP) dan kepuasan pengguna (KPG) tidak mempengaruhi efektivitas sistem informasi akademik pada kampus Universitas Mercu Buana. Sedangkan untuk kelompok pengguna Mahasiswa, kualitas sistem (KS), kualitas pelayanan (KP) dan kepuasan pengguna (KPG) mempengaruhi efektivitas sistem informasi akademik pada kampus Universitas Mercu Buana
Katakunci	AMOS, Universitas Mercu Buana, SEM, Sistem Informasi Akademik, Efektivitas
Penerbit	Universitas Mercu Buana

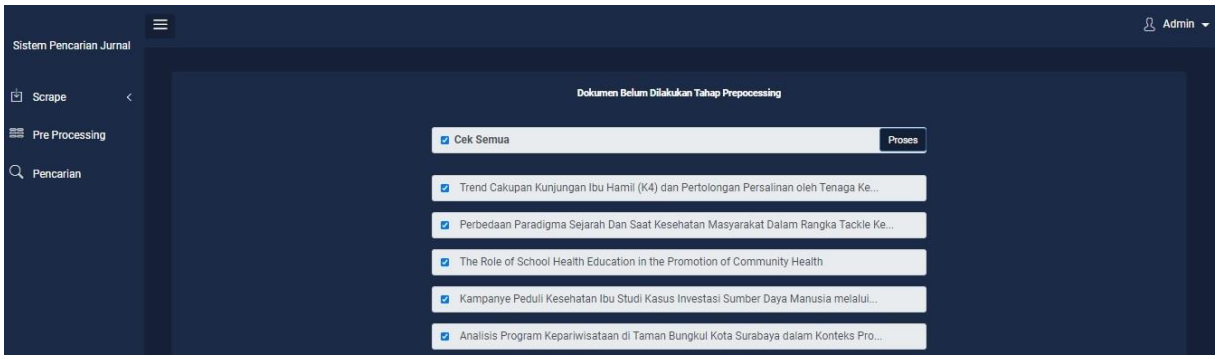
Building of Informatics, Technology and Science (BITS)
Volume 3, No 3, December 2021, Page 275−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i3.1058

Gambar 8. Hasil Proses Document Scraping



Gambar 9. Proses Klasifikasi Bahasa Dokumen

Proses berikutnya dilakukan proses stemming kata berdasarkan algoritma stemming yang dipilih berdasarkan klasifikasi artikel jurnal berdasarkan Bahasa sebagaimana yang ditunjukan pada Gambar 9. Pada penelitian ini penulis menggunakan library stemmer sastrawi untuk dokumen berbahahasa indonesia dan library snowball stemmer untuk dokumen berbahasa inggris. Pada Gambar 10 merupakan proses Praposes Documen pada aplikas



Gambar 10. Proses Praproses Documen Pada Aplikasi

Tabel 4. Hasil Proses Stemming Bahasa Indonesia

Hasil Proses Stemming Terhadap Data Set Artikel Jurnal
terap suatu sistem informasi organisasi kata hasil apabila sistem informasi sebut sukses laksana salah satu indikator sukses sistem informasi sistem informasi sebut efektif guna model sukses sistem informasi delone mclean teori praktek banyak dukung beberapa teliti ukur sukses sistem informasi organisasi teliti tuju laku kaji sukses sistem informasi akademik universitas mercu buana guna model sukses sistem informasi delone mclean data kumpul lalu survei mahasiswawi dosen karyawan guna sistem informasi akademik ambil sampel teliti guna kuesioner isi responden metode olah data analis model sama struktural structured equation modeling manfaat perangkat lunak amos hasil teliti dapat umum kualitas informasi ki guna p pengaruh efektivitas manfaat sistem informasi akademik kampus universitas mercu buana faktor faktor pengaruh kualitas sistem ks kualitas layan kp puas guna kpg hasil dapat dasar kelompok guna dosenkaryawan kualitas sistem ks kualitas layan kp puas guna kpg pengaruh efektivitas sistem informasi akademik kampus universitas mercu buana kelompok guna mahasiswa kualitas sistem ks kualitas layan kp puas guna kpg pengaruh efektivitas sistem informasi akademik kampus universitas mercu buana

Hasil Stemming pada Tabel 4, menunjukan bahwa kalimat sudah menjadi kumpulan term yang selanjutnya akan representasikan menggunakan Vector Space Model (VSM). Tahap pertama dari VSM adalah memberikan bobot

Building of Informatics, Technology and Science (BITS)
Volume 3, No 3, December 2021, Page 275−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i3.1058

sebuah term menggunakan TF-IDF menggunakan Persamaan 1. Hasil dari perhitungan TF-IDF ditunjukan pada Gambar 11

No	Term	Q	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D90	D91	D92	D93	D94	D95	D96	D97	D98	D99	D100	DF	IDF Log(n/df) + 1
1	sistem	1	17	6	10	17	9	4	10	14	7	6	5	13	6	6	0	0	0	0	0	0	0	0	14	0	0	30	1.527200119063
2	kaji	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	1	0	3	0	0	1	0	12	1.925140127735
3	efektivitas	0	4	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2.527200119063
4	informasi	0	15	13	9	13	9	5	8	14	10	11	12	13	10	10	0	0	2	0	0	0	0	0	0	0	0	22	1.6618986929604
5	akademik	0	7	0	2	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	2.4022613824547
6	universitas	0	6	0	2	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	5	2.3053513694466
7	mercu	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3.0043213737826
8	buana	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3.0043213737826
9	terap	0	1	0	0	0	0	0	0	1	0	0	1	0	3	2	0	0	0	0	0	0	0	0	1	0	0	9	2.0500788643433
2604	Sinapoy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	3.0043213737826

Gambar 11. Hasil Perhitungan TF-IDF

Setelah dilakukan proses perhitungan TF-IDF, Langkah berikutnya adalah membentuk representasi Vector Space Model yang menghasilkan beberapa vector diantaranya adalah vector dokumen dan vector query yang terdiri dari kumpulan term yang Menyusun setiap vektornya. Pada Gambar 12 merupakan hasil dari representasi Vector Space Model dan pada Gambar 13 merupakan hasil dari perhitungan Cosine Similarity yang dihitung menggunakan Persamaan 2.

No	Term	Q	D1	D2	D3	D4	D5	D6	D97	D98	D99	D100
1	sistem	1.527200119063	25.962402024071	9.1632007143779	15.27200119063	25.962402024071	13.744801071567	6.1088004762519	0	457.13867991853	0	0
2	kaji	0	3.85028025547	0	0	0	0	0	0	0	3.7061645114156	0
3	efektivitas	0	10.108800476252	0	0	0	0	0	0	0	0	0
4	informasi	0	24.928480394407	21.604683008486	14.957088236644	21.604683008486	14.957088236644	8.3094934648022	0	0	0	0
5	akademik	0	16.815829677183	0	4.8045227649094	0	0	0	0	0	0	0
6	universitas	0	13.83210821668	0	4.6107027388932	0	0	0	0	0	0	0
7	mercu	0	18.025928242696	0	0	0	0	0	0	0	0	0
8	buana	0	18.025928242696	0	0	0	0	0	0	0	0	0
9	terap	0	2.0500788643433	0	0	0	0	0	0	4.2028233500272	0	0
10	suatu	0	1.8581933381044	0	1.8581933381044	0	0	0	0	0	0	3.4528824817756
11	organisasi	0	4.3184466675368	0	0	2.1592233337684	0	0	0	0	0	0
2604	Sinapoy	0	0	0	0	0	0	0	0	0	0	9.0259469169672
SQRT(SUM(TF/IDF(Q,D))2) :		65.838670716717	39.75359642785	36.86680773742	61.746072978419	41.629343599201	25.070718532747	37.841220022506	38.608027643001	53.638333233541	49.226339046345	44.945738434863

Gambar 12. Representasi Perhitungan Vektor Space Model

PROSES 6 : HASIL PERSENTASE KEMIRIPAN TEKS

ID	Judul	Nilai Cosine
D1	Kajian Efektivitas Sistem Informasi Akademik Universitas Mercu Buana	0.39433363009072
D2	Pengaruh Teknologi Informasi terhadap Karakteristik Sistem Informasi Akuntansi Manajemen dan Dampaknya terhadap Kinerja Manajerial	0.23049991793845
D3	Sistem Informasi Berbasis Web Jurusan Sistem Informasi Fakultas Ilmu Komputer Universitas Sriwijaya	0.41424799509095
D4	Perencanaan Strategis Sistem Informasi Menggunakan Metode Ward dan Peppard di Direktorat Pengembangan Sistem Penyediaan Air Minum (Pspam)	0.42047049750265
D5	Sistem Informasi di Smk dan Upaya Peningkatan Kinerjanya	0.33017097756571
D6	Sistem Informasi Desa_2019	0.2436627601351
D7	Sinkronisasi Data User Antara Sistem Informasi Perpustakaan Dengan Sistem Informasi Akademik	0.40358109969887
D8	Efektivitas Implementasi Sistem Informasi Akuntansi Terintegrasi Pada PT. Ace Hardware Indonesia Tbk.	0.36689677894565
D9	Konsep Sistem Informasi	0.33123261589785
D10	Pengaruh Kualitas Sistem Informasi, Kualitas Informasi dan Perceived Usefulness pada Kepuasan Pengguna Akhir Software Akuntansi (Studi Empiris pada Hotel Berbintang di Provinsi Bali).	0.20137396077404
D100	Kearifan Lokal Masyarakat Adat Suku Moronene dalam Perlindungan dan Pengelolaan Lingkungan Hidup	0

Gambar 13. Hasil Perhitungan Cosine Similarity

Building of Informatics, Technology and Science (BITS)
Volume 3, No 3, December 2021, Page 275−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i3.1058

Gambar 14. Hasil Pencarian Pada Aplikasi

Untuk mengetahui hasil dari evaluasi system pada penelitian ini penulis melakukan percobaan pencarian sebanyak 6 query dengan proporsi 3 menggunakan term bahasa indonesia dan 3 term bahasa inggris dan direpresentasikan menggunakan confusion matrix untuk diketahui tingkat akurasi dihitung berdasarkan persamaan 3, dan presisi dihitung berdasarkan persamaan 4 sedangkan recall dihitung berdasarkan Persamaan 5. Tabel 5 menunjukan confusion matrix dari percobaan yang dilakukan, sedangkan Tabel 6 menunjukan hasil perbandingan kinerja dari query berbahasa Indonesia dan query berbahasa inggris.

Tabel 5. Tabel Confusion Matrix

Query	TP	TN	FP	FN	Akurasi	Presisi	Recall
ilmu	10	86	3	1	0,96	0,769230769	0,909090909
kesehatan	8	90	1	1	0,98	0,888888889	0,888888889
masyarakat	3	95	1	1	0,98	0,75	0,75
science	36	59	3	2	0,95	0,923076923	0,947368421
health	6	91	1	2	0,97	0,857142857	0,75
public	34	62	3	1	0,96	0,918918919	0,971428571

Tabel 6. Tabel Perbandingan Kinerja Jenis Query

Query	Akurasi	Presisi	Recall
Berbahasa Indonesia	0,973333333	0,802706553	0,849326599
Berbahasa Inggris	0,96	0,8997129	0,889598997
Total Rata-rata	0,966666667	0,851209726	0,869462798

4. KESIMPULAN

Berikut adalah kesimpulan dari analisa dan pembahasan dari penelitian tentang Penerapan Metode N-Gram dan Cosine Similarity Dalam Pencarian Pada Repositori Artikel Jurnal Publikasi, adapun kesimpulan yang diperoleh hasil evaluasi sistem dari penerapan metode N-Gram dan Cosine Similarity bekerja sangat baik untuk query berbahasa Indonesia dan query berbahasa inggris dengan total rata-rata akurasi sebesar 0,967. presisi sebesar 0,851 sedangkan rata recall didapatkan hasil 0,869. Kinerja sistem dalam menghitung nilai akurasi pada query berbahasa Indonesia didapatkan hasil lebih baik sebesar 0,973 daripada query berbahasa Indonesia sebesar 0,96. Dalam sistem yang dikembangkan query berbahasa inggris unggul pada tingkat presisi (0,097) dan recall (0,04) dari query berbahasa Indonesia.

REFERENCES

[1] A. Rachmat C., "Analisis Rancang Bangun Sistem Repositori Institusi Berbasis Metadata Dublin Core di UKDW Yogyakarta," *J. Ultim. InfoSys*, vol. 5, no. 2, pp. 65–74, 2014, doi: 10.31937/si.v5i2.267.

[2] I. G. Anugrah and H. Rosyid, "Penerapan Information Retrieval Menggunakan Pemodelan Topik Pada Deskripsi Portal Multimedia," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 2, no. 1, p. 48, 2019, doi: 10.32672/jnkti.v2i1.1057.

Building of Informatics, Technology and Science (BITS)
Volume 3, No 3, December 2021, Page 275−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v3i3.1058

[3] S. Azizurahman, Y. Firdaus, and A. A. Suryani, "Analisis Dan Implementasi Metode N-Gram Pada Information Retrieval," 2011.
[4] R. T. Wahyuni, D. Prastiyanto, and E. Supraptono, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi," *J. Tek. Elektro*, vol. 9, no. 1, pp. 18–23, 2017, doi: 10.15294/jte.v9i1.10955.
[5] S. Sugiyamto, B. Surarso, and A. Sugiharto, "Analisa Performa Metode Cosine Dan Jacard Pada Pengujian Kesamaan Dokumen," *J. Masy. Inform.*, vol. 5, no. 10, 2014, doi: 10.14710/jmasif.5.10.1-8.
[6] B. Zaman, E. Hariyanti, and E. Purwanti, "Sistem Deteksi Bahasa pada Dokumen menggunakan N-Gram," *Multinetics*, vol. 1, no. 2, p. 21, 2015, doi: 10.32722/vol1.no2.2015.pp21-26.
[7] D. L. Khuseri Andesa, "Implementasi Vector Space Model Untuk Meningkatkan," *Semin. Nas. Inform. 2015*, no. May 2013, pp. 8–15, 2015.
[8] D. D. A. Yani, H. S. Pratiwi, and H. Muhardi, "Implementasi Web Scraping untuk Pengambilan Data pada Situs Marketplace," *J. Sist. dan Teknol. Inf.*, vol. 7, no. 4, p. 257, 2019, doi: 10.26418/justin.v7i4.30930.
[9] I. Fakhruddin and I. G. Anugrah, "Implementation of Winnowing Algorithm and Simple Additive Weighting SAW for Publication Reference Journal Search System," *J. Dev. Res.*, vol. 5, no. 2, pp. 61–72, 2021, doi: 10.28926/jdr.v5i2.141.
[10] D. N. Chandra, G. Indrawan, and I. N. Sukajaya, "Klasifikasi Berita Lokal Radar Malang Menggunakan Metode Naïve Bayes Dengan Fitur N-Gram," *J. Ilm. Teknol. Inf. Asia*, vol. 10, no. 1, pp. 11–19, 2016.
[11] D. S. Indraloka and B. Santosa, "Penerapan Text Mining untuk Melakukan Clustering Data Tweet Shopee Indonesia," *J. Sains dan Seni ITS*, vol. 6, no. 2, pp. 6–11, 2017, doi: 10.12962/j23373520.v6i2.24419.
[12] M. S. Anwar, I. M. I. Subroto, and S. Mulyono, "Sistem Pencarian E-Journal Menggunakan Metode Stopword Removal Dan Stemming," *Pros. Konf. Ilm. Mhs. UNISSULA 2*, pp. 58–70, 2019, [Online]. Available: http://lppm-unissula.com/jurnal.unissula.ac.id/index.php/kimueng/article/viewFile/8420/3887.
[13] P. F. Ariyani, A. Rahmala, and N. Juliasari, "Implementasi Metode Stemming Tala Dan Fungsi Jaccard Pada Aplikasi Katalog Perpustakaan," *Semin. Nas. Inov. dan Apl. Teknol. di Ind. 2019*, pp. 128–133, 2019.
[14] I. M. A. Agastya, "Pengaruh Stemmer Bahasa Indonesia Terhadap Peforma Analisis Sentimen Terjemahan Ulasan Film," *J. Tekno Kompak*, vol. 12, no. 1, p. 18, 2018, doi: 10.33365/jtk.v12i1.70.
[15] I. Made Suwija Putra, N. Putu Ayu Widiari, and I. Wayan Gunaya, "Implementasi Generalized Vector Space Model (GVSM) dalam Pencarian Buku di Perpustakaan," *J. Ilm. Merpati (Menara Penelit. Akad. Teknol. Informasi)*, vol. 7, no. 1, p. 86, 2019, doi: 10.24843/jim.2019.v07.i01.p10.
[16] A. Apriani, H. Zakiyudin, and K. Marzuki, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF System Penerimaan Mahasiswa Baru pada Kampus Swasta," *J. Bumigora Inf. Technol.*, vol. 3, no. 1, pp. 19–27, 2021, doi: 10.30812/bite.v3i1.1110.
[17] I. W. Saputro and B. W. Sari, "Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa," *Creat. Inf. Technol. J.*, vol. 6, no. 1, p. 1, 2020, doi: 10.24076/citec.2019v6i1.178.