PENINGKATAN HASIL SISTEM TEMU KEMBALI INFORMASI BERBASIS PADA KATA MAJEMUK MENGGUNAKAN JACCARD SIMILARITY

Dwi Agus Diartono, Isworo Nugroho, Jeffri Alfa Razaq, Jati Sasongko Wibowo*
Fakultas Teknologi Informasi dan Industri, Universitas Stikubank
Email: *jatisw@edu.unisbank.ac.id

ABSTRAK

Kata majemuk gabungan morfem dasar yang kesemuanya berstatus kata dengan pola fonetik, gramatikal, dan semantik tertentu sesuai dengan kaidah bahasa yang bersangkutan. Pola khusus ini membedakannya dari frasa atau kombinasi kata—kombinasi morfem yang bukan kata majemuk. Pembatasan ini berlaku untuk sejumlah dimensi, dan persamaan kosinus paling sering digunakan dalam ruang positif berdimensi tinggi. Misalnya, dalam pencarian informasi dan penambangan teks, setiap istilah secara tidak langsung diberi dimensi yang berbeda, dan dokumen diberi label sebagai vektor, di mana nilai di setiap dimensi sesuai dengan berapa kali istilah tersebut muncul dalam dokumen. Kesamaan cosine kemudian memberikan ukuran yang berguna tentang seberapa mirip dua dokumen dalam hal topik. Pada penelitian ini penggunaan kata majemuk sebagai kata kunci dalam pencarian dokumen menghasilkan dokumen yang lebih tepat. Data yang digunakan sebanyak 15 file. Data yang mengandung kata majemuk sebanyak 10 file. Data 5 dokumen tanpa kata majemuk. Hasil pencarian dengan menggunakan algoritma kesamaan menghasilkan nilai kesamaan yang tinggi, namun 33,3% kata majemuk tidak ada. Dan hasil pencarian dengan nilai kemiripan lebih rendah tetapi mengandung kata majemuk sebesar 66,3%. Sehingga dengan adanya penambahan data berupa kata majemuk di dalam pencarian membuat hasilnya lebih baik.

Kata kunci: sistem temu kembali informasi, kata majemuk, cosine similarity

1. PENDAHULUAN

Sistem pencarian informasi digunakan untuk secara otomatis mengambil informasi yang terkait dengan kebutuhan pengguna dari kumpulan informasi. Aplikasi umum dari sistem pencarian informasi adalah mesin pencari atau mesin pencari yang ditemukan di Internet. Pengguna dapat mencari halaman web yang mereka butuhkan melalui mesin.

Ukuran efektivitas pencarian ditentukan oleh presisi dan daya ingat. Presisi adalah rasio jumlah dokumen relevan yang ditemukan oleh mesin pencari terhadap jumlah total dokumen yang ditemukan. Presisi menunjukkan seberapa bagus set jawaban, tetapi tidak melihat jumlah total dokumen yang relevan dalam set dokumen. Recall adalah rasio jumlah dokumen relevan yang ditemukan kembali dengan jumlah total dokumen yang dianggap relevan.

Dalam pencarian informasi, mendapatkan dokumen yang relevan saja tidak cukup. Tujuan yang harus dicapai adalah bagaimana mendapatkan file yang relevan, bukan file yang tidak relevan. Tujuan lainnya adalah bagaimana mengurutkan dokumen yang diperoleh, menampilkannya secara berurutan dari yang relevansinya lebih tinggi ke yang relevansinya lebih rendah. Menyusun dokumen secara berurutan disebut pengurutan dokumen. Model ruang vektor dan model probabilistik adalah dua cara untuk melakukannya.

Model ruang vektor dan model probabilistik adalah model yang menggunakan bobot kata dan peringkat dokumen. Hasil temu kembali yang diperoleh dari model tersebut adalah mengurutkan dokumen yang dianggap paling relevan dengan query. Ada beberapa cara atau

pendekatan pembobotan kata dalam pendekatan TF-IDF, yaitu dengan skema pembobotan query dan dokumen.

Dalam model ruang vektor, dokumen dan kueri direpresentasikan sebagai vektor yang diurutkan berdasarkan suku terindeks dalam ruang vektor, yang kemudian dimodelkan dengan persamaan geometris. Pada saat yang sama, model probabilistik membuat asumsi tentang distribusi istilah dalam dokumen yang relevan dan tidak relevan untuk memperkirakan kemungkinan relevansinya dengan kueri.

2. TINJAUAN PUSTAKA

2.1. Kata Majemuk

Kata majemuk adalah gabungan morfem dasar yang kesemuanya berstatus kata dengan pola fonetik, gramatikal, dan semantik tertentu sesuai dengan kaidah bahasa yang bersangkutan [1][2]. Pola khusus ini membedakannya dari frasa atau kombinasi kata—kombinasi morfem yang bukan kata majemuk [3]. Misalnya, dalam bahasa Indonesia, aju hijau adalah frasa dan kamar mandi adalah kata majemuk; dalam bahasa Inggris, black bird adalah frasa dan blackbird adalah kata majemuk [4].

Kata majemuk dibentuk oleh proses pemajemukan atau kombinasi, yang merupakan proses leksikal sedangkan frase dibentuk oleh proses sintaksis [3]. Kata majemuk bahasa Indonesia memiliki ciri-ciri sebagai berikut: (1) Non-insertable, yaitu tidak ada yang dapat disisipkan antar komponen majemuk; (2) Non-extensible, yaitu setiap komponen bahan komposit hanya dapat ditambahkan pada satu waktu. ; (3) Irreversibility, yaitu unsur-unsur komposit tidak dapat dipertukarkan [1].

2.2. Cosine Similarity

Kesamaan cosinus adalah ukuran kesamaan antara dua vektor bukan nol dari ruang produk dalam, yang mengukur cosinus sudut di antara keduanya [5][6]. Kosinus 0° adalah 1, dan setiap sudut dalam interval radian (0, π) kurang dari 1. Jadi, ini adalah penilaian arah dan bukan besaran: dua vektor dalam arah yang sama memiliki persamaan kosinus 1, dua vektor pada 90° relatif satu sama lain memiliki kesamaan 0, dan dua vektor dengan diameter berlawanan memiliki kesamaan -1 terlepas dari ukurannya. Di {\displaystyle[0,1]}[0,1]. Namanya berasal dari istilah "arah kosinus": dalam hal ini, vektor satuan secara maksimal "mirip" jika paralel, dan maksimal "berbeda" jika ortogonal (tegak lurus). Ini mirip dengan cosinus, yaitu satu (nilai maksimum) ketika sudut cenderung ke nol ketika ruas garis tegak lurus dan nol (tidak berkorelasi) [7].

Pembatasan ini berlaku untuk sejumlah dimensi, dan persamaan kosinus paling sering digunakan dalam ruang positif berdimensi tinggi. Misalnya, dalam pencarian informasi dan penambangan teks, setiap istilah secara tidak langsung diberi dimensi yang berbeda, dan dokumen diberi label sebagai vektor, di mana nilai di setiap dimensi sesuai dengan berapa kali istilah tersebut muncul dalam dokumen. Kesamaan cosine kemudian memberikan ukuran yang berguna tentang seberapa mirip dua dokumen dalam hal topik.[8]

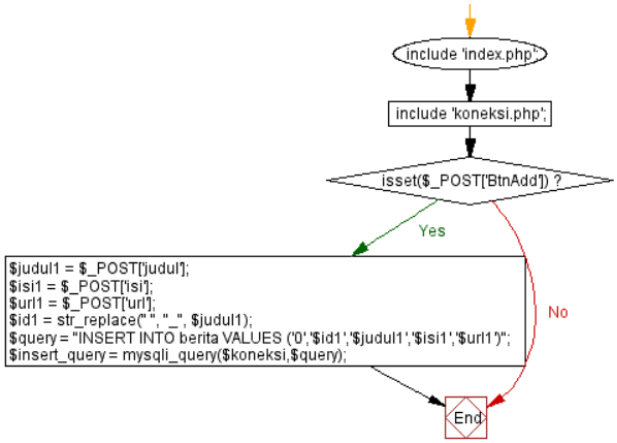
Teknik ini juga digunakan untuk mengukur kohesi di dalam cluster di bidang data mining. Istilah cosine distance sering digunakan untuk komplemen dalam ruang positif, yaitu: {\ displaystyle D_ {C} (A, B) = 1-S_ {C} (A, B),}{\ displaystyle D_ {C} (A, B) = 1-S_ {C} (A, B),} dimana {\ displaystyle D_ {C}}D_C adalah jarak cosinus dan {\ displaystyle S_ {C}}S_ {C}adalah kesamaan cosinus. Penting untuk dicatat, bagaimanapun, bahwa ini bukan metrik jarak yang tepat karena tidak memiliki properti ketidaksetaraan segitiga - atau, lebih formal, ketidaksetaraan Schwarz - dan itu melanggar aksioma kebetulan; untuk memperbaiki properti ketimpangan segitiga sambil mempertahankan urutan yang sama, perlu untuk

mengkonversi ke jarak sudut. Salah satu keuntungan dari kesamaan cosinus adalah kompleksitasnya yang rendah, terutama untuk vektor jarang : hanya dimensi non-nol yang perlu dipertimbangkan [9].

3. METODE PENELITIAN

3.1. Input Data

Input data pada penelitian ini merupakan langkah paling awal untuk mendapatkan data. Data di masukkan melalui form yang sudah disediakan dengan tiga field isian diantaranya judul dokumen, isi dari dokumen, dan url sumber dari dokumen. Data disimpan dalam database dengan nama majemuk dan tabel dengan nama tabel berita. Proses input data melalui form dapat digambarkan melalui flowchart seperti pada gambar 1.



Gambar 1 Flowchart Form Input Data Dokumen

Data yang digunakan untuk penelitian ini sebanyak dua puluh data record.. Data diambilkan dari berbagai sumber yang berasal dari pencarian di google. Pencarian sample data melalui google dengan kata kunci kursi dan roda yang tiap kata dipisah. Kata kunci yang lain menggunakan kata majemuk kursi roda yang tiap katanya tidak dipisah..

Hasil input data dokumen dapat dilihat pada tabel 1 yang berisi id dokumen, judul dokumen, isi dokumen dan url sumber dari dokumen. Jumlah data sebanyak dua puluh dokumen. Data masih dalam bentuk asli dari sumber pencarian di mesin pencari google, dimana data masih terdapat huruf besar dan huruf kecil, mempunyai tanda baca, symbol dan angka.

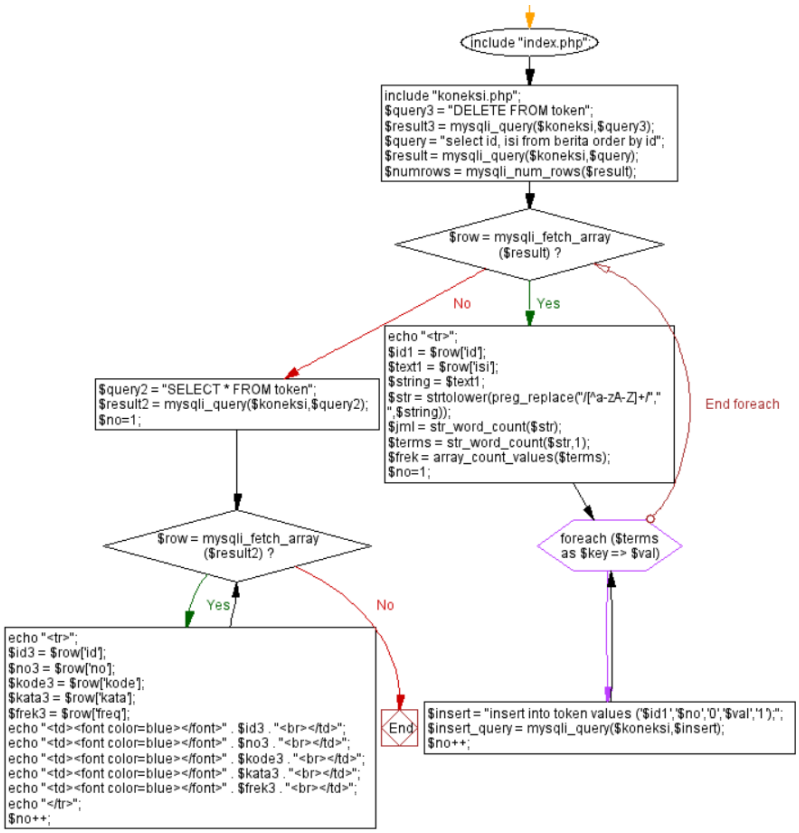
Tabel 1 Hasil Input Data Dokumen

ID	Judul	Isi	Url
1	Roda kursi kantor 1 set (isi 5 buah) | Shopee Indonesia	Cocok untuk semua jenis kursi kerja Tidak perlu kursi baru cukup ganti roda serasa seperti baru kembali #rodakursi #roda #murahmeriah #surabaya #ootd ...	https://shopee.co.id/Roda-kursi-kantor-1-set-(isi-5-buah)-i.5916531.46057265
2	RODA ETALASE BAUT BULAT KARET SKK / RODA KURSI ...	RODA SKK 1 SET ISI 4 PCS tipe baut diameter roda:5 cm bahan roda:karet Dapat digunakan pada etalase, kursi sofa,meja,kaki kulkas dan lain lain Kwalitas ...	https://shopee.co.id/RODA-ETALASE-BAUT-BULAT-KARET-SKK-RODA-KURSI-LEMARI-RAK-ISI-4-PCS-i.10619021.866295462
3	Harga Kursi Kantor Terbaru di Indonesia September 2020	Kursi Kerja Tanpa Roda: Beberapa pilihan kursi kerja tidak beroda yang bisa Anda beli secara online di iprice Indonesia adalah Savello Office Chair Trinity DX ...	https://iprice.co.id/ruang-kerja/kursi/
4	Blog - Roda Adalah “Sumber Masalah” Pada Kursi Kantor, Ini ...	Oct 16, 2019 - Kursi kantor Anda terasa sulit digerakkan? Bisa jadi masalahnya ada pada roda kursinya, lho! Gudang Furniture akan memaparkan beberapa ...	https://gudangfurniture.com/blog/article/roda-adalah-sumber-masalah-pada-kursi-kantor-ini-lho-cara-merawatnya
5	Jual Galena Kursi Pispot Aluminum Dengan Roda Fs 699 L ...	Rangka kursi terbuat dari material aluminium dengan pijakan kaki yang dapat dilipat serta roda agar lebih mudah dipindahkan. Kursi Pispot ini juga dilengkapi ...	https://www.ruparupa.com/galena-kursi-pispot-aluminum-dengan-roda-fs699l.html

6	Jual Kursi Kantor Ergosit Terbaru - Harga Promo | Blibli.com	Jual Kursi Kantor Ergosit Terbaru - Daftar Harga Terupdate & Terbaru, Harga Promo & Diskon, ... Ergosit Castor Yc 6381 Part atau Roda Kursi Kantor [5 pcs].	https://www.blibli.com/brand/ergosit
7	Jual Kursi Kantor Minimalis Terlengkap | IKEA Indonesia	Lihat katalog kursi kantor IKEA untuk keperluan bisnis Anda. ... Kursi konferensi dengan roda, veneer kayu ash diwarnai hitam/Gunnared abu-abu tua. Rp 3.299.	https://www.ikea.co.id/in/produk/perabotan-kantor/kursi-kantor
8	jual kursi kantor roda - Home Furniture Jepara	Home / Products tagged “jual kursi kantor roda”. Filter. Showing the single result. Default sorting, Sort by popularity, Sort by average rating, Sort by latest, Sort by ...	https://www.homefurniturejepara.com/product-tag/jual-kursi-kantor-roda/
9	Chitose NA - PT. Chitose Internasional Tbk	Saat ini perusahaan kami menggunakan kursi chitose tipe NA. Yang mau ditanyakan, bagaimana dan di mana bisa dibeli castor/roda untuk kursi tipe tersebut?	https://www.chitose-indonesia.com/produk/na/
10	Kursi Kamar Mandi ini kami Jual dengan Harga Ekonomis	Toko penjual alat bantu kursi kamar mandi pasien untuk BAB dengan harga yang terjangkau. spesifikasi kursi kamar mandi ini secara detail ada di website ini.	https://www.alatkesehatan.id/toko/kursi-kamar-mandi-13122/

3.2.Tokenization

Proses tokenization merupakan salah satu proses penting dalam sistem pencarian dokumen, yaitu memisahkan kata berdasarkan spasi dari database dokumen yang isinya berupa kalimat. Sehingga kalimat dalam dokumen yang terdiri dari kata, maka akan dipisahkan per kata dan simpan kembali ke dalam database baru. Apabila dalam satu kalimat terdiri dari seribu kata, maka kalimat tersebut akan dipisah per kata sejumlah seribu kata. Maka akan di dapatkan database berupa data yang berisi kata saja. Selain kata yang disimpan dalam database, disimpan juga id untuk menandakan kata yang disimpan berasal dari dokumen yang mana. Ada juda data yang ikut disimpan dalam proses tokenization ini yaitu frekuensi jumlah kata dalam satu dokumen, tetapi dalam penelitian ini nilai frekuensi dimasukkan nilai satu semua. Sehingga untuk menghitung jumlah kata tinggal dijumlahkan dari daftar frekuensi yang telah disimpan sebelumnya.



Gambar 2 Flowchart Tokenization

Proses tokenization pada penelitian ini menggunakan beberapa perintah di bahasa pemrograman php. Perintah utama yang digunakan diantaranya perintah mysqli_fetch_array yang berfungsi untuk menyimpan data hasil query ke dalam sebuah variable. Terdapat perintah str_word_count yang berfungsi untuk menghitung data array beserta dengan indexnya. Dari data array kemudian data per index atau secara urut diproses dipisahkan dan disimpan ke dalam tabel.

Hasil proses tokenization terhadap data dokumen dapat dilihat pada tabel 2 yang berisi id dokumen, no dokumen, kode dokumen, kata hasil token, frekuensi jumlah kata. Id dokumen menunjukkan no identitas dokumen. No dokumen menunjukkan nomor urut kata pada sebuah dokumen. Kode menunjukkan terdapat kata majemuk atau tidak. Kata merupakan hasil token. Frekuensi menunjukkan nilai kata pada sebuah dokumen.

Tabel 2 Hasil Proses Tokenization

Id	No	Kode	Kata	Freq
1	1	0	cocok	1
1	2	0	untuk	1
1	3	0	semua	1
1	4	0	jenis	1
1	5	0	kursi	1
1	6	0	kerja	1
1	7	0	tidak	1
1	8	0	perlu	1
1	9	0	kursi	1
1	10	0	baru	1

3.3.Proses Kata Majemuk

Proses majemuk tahap 1 bertujuan untuk mendapatkan kata majemuk dalam data hasil token. Dengan menggunakan urutan kata pada hasil token, kata pertama akan dipasangkan dengan kata kedua, kata kedua akan dipasangkan dengan kata ketiga dan seterusnya, maka setiap pasangan akan di dapatkan dua kata. Pasangan dua kata yang terbentuk akan di cek berdasarkan daftar kamus kata majemuk yang sudah ada, apakah pasangan kata tersebut termasuk kata majemuk atau bukan.

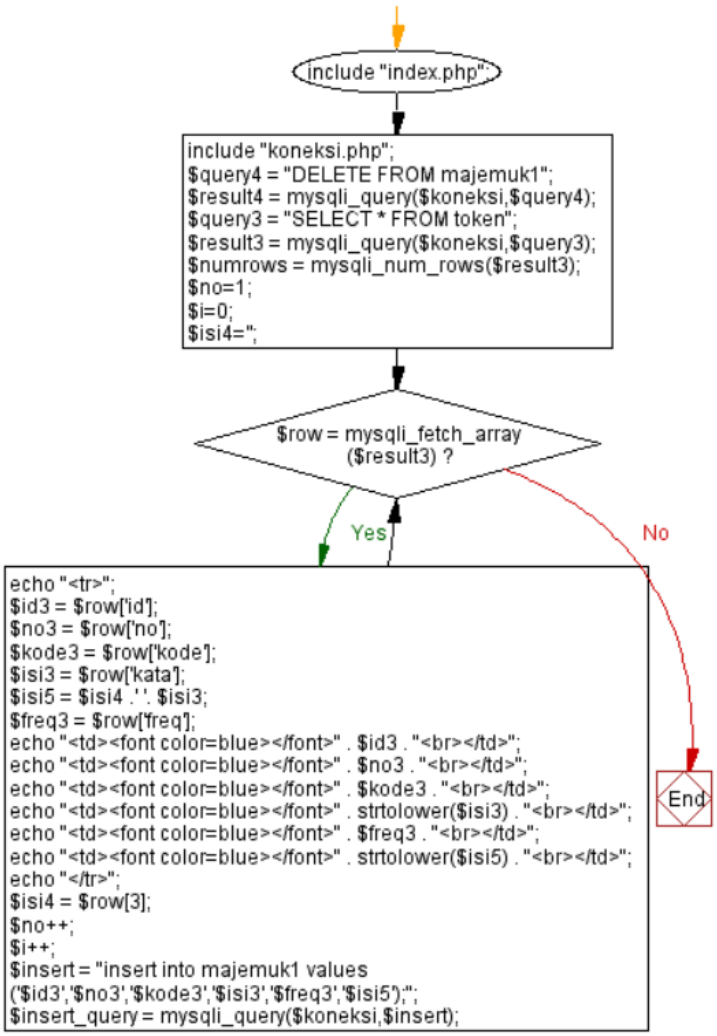
Proses majemuk 1 secara pemrograman menggunakan data array untuk mendapatkan setiap kata dari data hasil token. Data array secara index akan dipasang-pasangkan, dalam arti bahwa data array index pertama akan dipasangkan dengan data array index kedua, data array index kedua akan dipasangkan dengan data array index ketiga, dan seterusnya. Data array yang sudah dipasangkan akan disimpan secara tersendiri.

Tabel 3 Hasil Proses Cek Kata Majemuk Tahap 1

Id	No	Kode	Kata	Freq	Majemuk
1	1	0	cocok	1	cocok
1	2	0	untuk	1	cocok untuk
1	3	0	semua	1	untuk semua
1	4	0	jenis	1	semua jenis
1	5	0	kursi	1	jenis kursi
1	6	0	kerja	1	kursi kerja
1	7	0	tidak	1	kerja tidak
1	8	0	perlu	1	tidak perlu
1	9	0	kursi	1	perlu kursi
1	10	0	baru	1	kursi baru

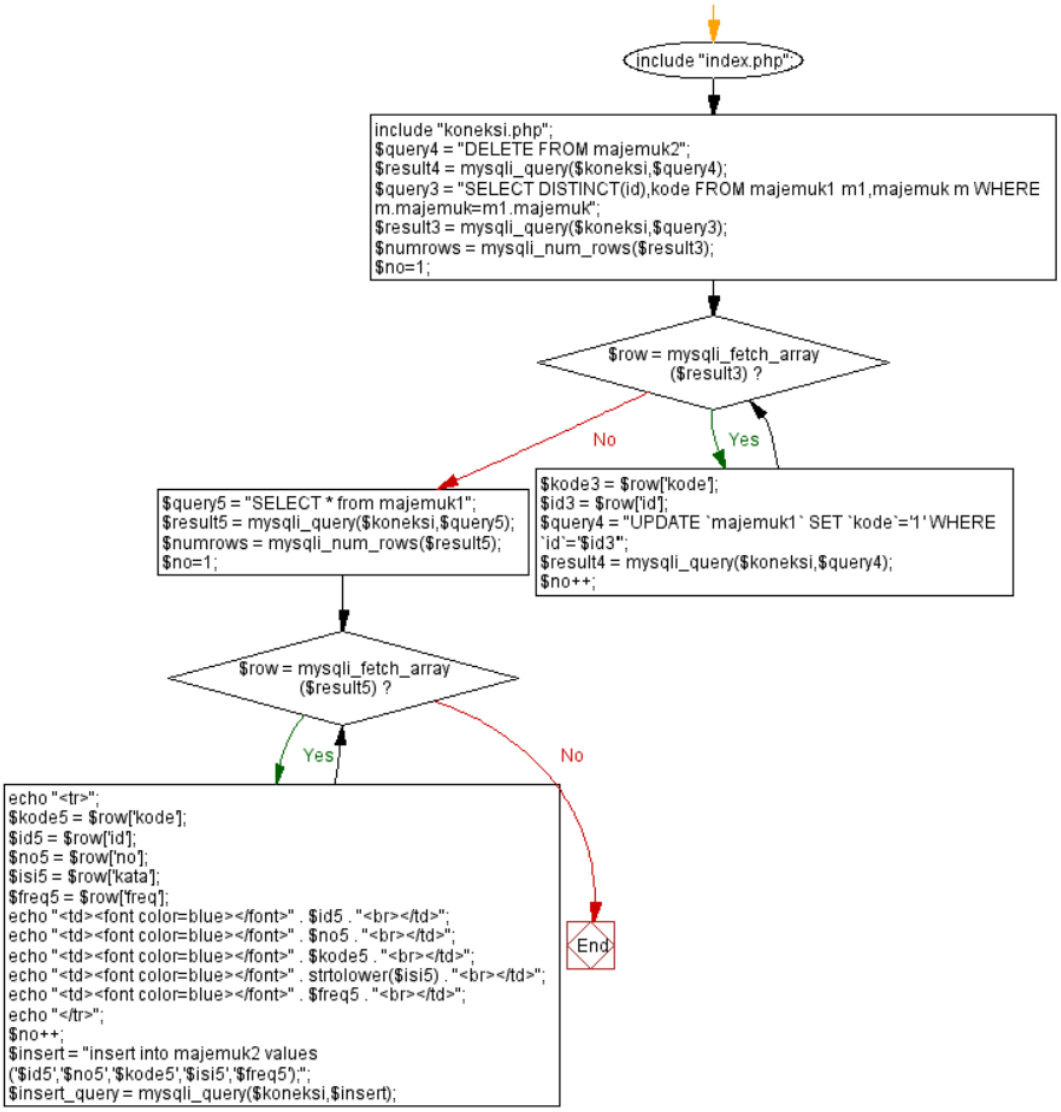
Hasil dari proses tokenization dari data dokumen dapat dilihat pada tabel 3 yang berisi id dokumen, no dokumen, kode dokumen, kata hasil token, frekuensi jumlah kata. Id dokumen menunjukkan no identitas dokumen. No dokumen menunjukkan nomor urut kata pada sebuah dokumen. Kode menunjukkan terdapat kata majemuk atau tidak. Kata merupakan hasil token.

Frekuensi menunjukkan nilai kata pada sebuah dokumen. Majemuk menunjukkan pasangan kata yang terdiri dari dua kata untuk setiap kata yang berurutan.



Gambar 3 Flowchart Proses Cek Kata Majemuk Tahap 1

Proses majemuk tahap dua bertujuan untuk mendeteksi keberadaan kata majemuk dari kata yang sudah dipasang-pasangkan dalam database, dengan membandingkan daftar kamus kata majemuk yang sudah tersedia juga dalam database. Apabila kata pasangan yang terbentuk setelah terdeteksi merupakan kata majemuk maka dalam database diberitan tanda atau kode satu. Dan apabila kata pasangan yang terbentuk tidak terdeteksi sebagai kata majemuk maka dalam database diberikan tanda atau kode nol. Kode atau tanda yang diberikan untuk semua token pada id dokumen yang sama, yang berarti menandakan bahwa dalam satu dokumen terdapat kata majemuk.



Gambar 4 Flowchart Proses Cek Kata Majemuk Tahap 2

Proses cek kata majemuk tahap dua digunakan untuk memberikan tanda pada data dokumen yang terdapat kata majemuk. Data dokumen yang memiliki kata majemuk diberikan tanda pada field kode. Field kode akan diberikan nilai nol apabila pada data dokumen tidak terdapat kata majemuk. Field kode akan diberikan nilai satu apabila pada data dokumen terdapat kata majemuk. Deteksi kata majemuk dengan membandingkan data token dengan data kamus kata majemuk menggunakan perintah query select. Apabila sesuai dengan kata majemuk maka data akan dilakukan update dengan menggunakan perintah sql update. Sehingga akan terlihat bahwa semua kode yang mempunyai nilai satu berarti terdapat kata majemuk. Sedangkan kode yang mempunyai nilai nol maka dalam datanya tidak terdapat kata majemuk.

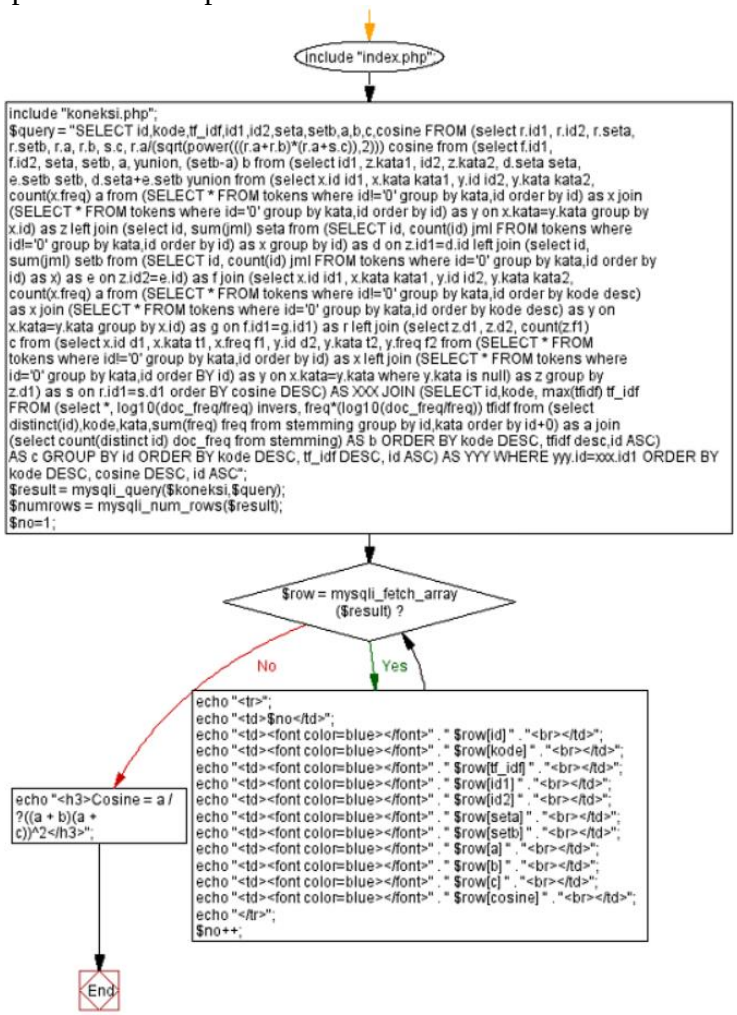
Hasil dari proses tokenization dari data dokumen dapat dilihat pada tabel 4 yang berisi id dokumen, no dokumen, kode dokumen, kata hasil token, frekuensi jumlah kata. Id dokumen menunjukkan no identitas dokumen. No dokumen menunjukkan nomor urut kata pada sebuah dokumen. Kode menunjukkan terdapat kata majemuk atau tidak. Kata merupakan hasil token. Frekuensi menunjukkan nilai kata pada sebuah dokumen. Majemuk menunjukkan pasangan kata yang terdiri dari dua kata dari urutan index kata pada dokumen.

Tabel 4 Hasil Proses Cek Kata Majemuk Tahap 2

Id	No	Kode	Kata	Freq	Majemuk
1	1	0	cocok	1	cocok
1	2	0	untuk	1	cocok untuk
1	3	0	semua	1	untuk semua
1	4	0	jenis	1	semua jenis
1	5	0	kursi	1	jenis kursi
1	6	0	kerja	1	kursi kerja
1	7	0	tidak	1	kerja tidak
1	8	0	perlu	1	tidak perlu
1	9	0	kursi	1	perlu kursi
1	10	0	baru	1	kursi baru

3.4.Proses Cosine Similarity

Proses cosine similarity merupakan proses untuk menentukan kemiripan dari kata kunci yang diinputkan oleh user pada mesin pencari dengan data dokumen yang terdapat dalam basis data mesin pencari. Kemiripan dari kata kunci dengan dokumen yang ada di nilai dari angka nol hingga angka satu, dimana angka nol berarti bahwa tidak mempunyai kemiripan sama sekali, sedangkan angka satu menginformasikan bahwa kemiripan dokumen dianggap sama persis. Sehingga semakin mendekati angka nol berarti tidak mirip, dan semakin mendekati angka satu dianggap semakin mirip.



Gambar 5 Flowchart Proses Cosine Similarity

Pada proses cosine similarity untuk pengembangan sistemnya menggunakan sql yang dipadukan dengan php. Perintah cosine similarity sendiri murni menggunakan bahasa sql yang cukup panjang apabila dituliskan sebagai satu queri. Dapat dilihat pada gambar 5 bahwa sebuah queri untuk memproses cosine similarity dibutuhkan satu query dengan jumlah baris sampai tiga puluhan. Sampai sebanyak itu dikarenakan semua proses dari mencari nilai freq, nilai document frequency, nilai invers document frequency masuk dalam satu queri dengan queri cosine similarity.

Hasil dari cosine similarity dari data dokumen dapat dilihat pada tabel 5 yang berisi id dokumen, kode dokumen, term frequency invers document frequency , set a, set b, a, b, c, dan cosine. No menunjukkan no urut record. Id dokumen menunjukkan no identitas dokumen. Kode menunjukkan terdapat kata majemuk atau tidak. Tfidf menunjukkan nilai bobot dari setiap dokumen. Id1 dan Id2 menunjukkan no identitas dari kata kunci dan data dokumen. Set a menunjukkan himpunan a yang mewakili jumlah kata dalam sebuah dokumen. Set b menunjukkan himpunan b yang mewakili jumlah kata dalam kata kunci. Merupakan irisan dari himpunan a dan himpunan b. b merupakan kata kunci diluar irisan a. c merupakan data dokumen diluar irisan a. Cosine menunjukkan nilai kemiripan antara kata kunci dengan data dokumen. Semakin mendekati nilai 0 semakin tidak mirip, semakin mendekati nilai 1 semakin mirip.

Tabel 5 Hasil Proses Cosine Similarity

No	Id	Kode	TfIdf	Id1	Id2	Set A	Set B	a	b	c	Cosine
1	13	1	2.7958800173440754	13	0	11	2	2	0	9	0.09090909090909091
2	19	1	1.3010299956639813	19	0	12	2	2	0	10	0.08333333333333333
3	12	1	1.3010299956639813	12	0	15	2	2	0	13	0.06666666666666667
4	15	1	1.3010299956639813	15	0	15	2	2	0	13	0.06666666666666667
5	18	1	2	18	0	15	2	2	0	13	0.06666666666666667
6	11	1	1.3010299956639813	11	0	16	2	2	0	14	0.0625
7	16	1	2	16	0	16	2	2	0	14	0.0625
8	20	1	2	20	0	16	2	2	0	14	0.0625
9	14	1	2	14	0	17	2	2	0	15	0.058823529411764705
10	17	1	2	17	0	17	2	2	0	15	0.058823529411764705
11	9	0	2	9	0	8	2	2	0	6	0.125
12	4	0	2	4	0	10	2	2	0	8	0.1
13	1	0	2	1	0	11	2	2	0	9	0.09090909090909091
14	3	0	2	3	0	13	2	2	0	11	0.07692307692307693
15	5	0	2	5	0	13	2	2	0	11	0.07692307692307693

4. KESIMPULAN

Pada penelitian ini penggunaan kata majemuk untuk kata kunci dalam pencarian dokumen menghasilkan dokumen yang lebih tepat. Data yang digunakan sejumlah 15 dokumen. Data yang mempunyai kata majemuk sebanyak 10 dokumen. Data yang tidak memiliki kata majemuk 5 dokumen. Hasil pencarian yang menggunakan algoritma similarity menghasilkan nilai kemiripan tinggi tetapi tidak mempunyai kata majemuk sejumlah 33.3%. Sedangkan hasil pencarian yang mempunyai nilai kemiripan lebih rendah tetapi mempunyai kata majemuk sebanyak 66.3%. Sehingga dengan adanya penambahan data berupa kata majemuk di dalam pencarian membuat hasilnya lebih baik dibandingkan tidak mempunyai data kata majemuk.

DAFTAR PUSTAKA

[1] Kebol, Y. J. (2022). Kata Majemuk Bahasa Manggarai (Suatu Kajian Teoretis). Prolitera: Jurnal Penelitian Pendidikan, Bahasa, Sastra, Dan Budaya, 5(1), 60-75. https://doi.org/10.36928/jpro.v5i1.1363

[2] Faradilla, N. A. N. ., Wulandari, R. A. ., Putantri, W. ., & Ulya , C. . (2021). Analisis Kesalahan Berbahasa Bidang Morfologi Pada Portal Berita Online Esensinews.Com. Jurnal Review Pendidikan Dan Pengajaran (JRPP), 4(2), 344–352. https://doi.org/10.31004/jrpp.v4i2.3243

[3] Yumni, N. Z., Chaerunnissa, Hadana, I. N. ., Arimbi, S. D. ., & Utomo, A. P. Y. . (2022). Analisis Kalimat Majemuk dalam Novelet Wayang Tembang Cinta Para Dewi pada Bab "Dendam Abadi Seorang Dewi" Karya Naning Pranoto. JURNAL RISET RUMPUN ILMU BAHASA, 1(1), 71–87. https://doi.org/10.55606/jurribah.v1i1.124

[4] Aisyiah Syiam Octavianti, Fika Uswatun, Sefiyan Eza Nur Hidayat, & Asep Purwo Yudi Utomo. (2022). Analisis Penggunaan Frasa Verba pada Surat Kabar Suara Merdeka yang Berjudul "Kurikulum Ruh Pembelajaran Tingkat Paling Dasar hingga Bangku Kuliah": Analysis of the Use of Verb Phrases in Suara Merdeka Newspaper entitled "The Curriculum of the Most Basic Level of Learning Spirit to Lecturers". Jurnal Pendidikan Dan Sastra Inggris, 2(1), 77–85. https://doi.org/10.55606/jupensi.v2i1.190

[5] Lahitani, A. R. (2022). Automated Essay Scoring menggunakan Cosine Similarity pada Penilaian Esai Multi Soal. Jurnal Kajian Ilmiah, 22(2), 107–118. https://doi.org/10.31599/jki.v22i2.1121

[6] Arif Zulvian, S., Prihandan, K., & Ridha, A. A. (2021). Perbandingan Metode MSD dan Cosine Similarity pada Sistem Rekomendasi Item-Based Collaborative Filtering. INTECOMS: Journal of Information Technology and Computer Science, 4(2), 340 - 347. https://doi.org/https://doi.org/10.31539/intecoms.v4i2.2781

[7] N. W. Utami and I. G. J. . Eka Putra, "Text Minig Clustering Untuk Pengelompokan Topik Dokumen Penelitian Menggunakan Algoritma K-Means Dengan Cosine Similarity", JINTEKS, vol. 4, no. 3, pp. 255-259, Aug. 2022.

[8] Suarnata, I Gede; Sukarsa, I Made; Wibawa, Kadek Suar. Pencocokan Menu Berbasis Keywords pada Chatbot dengan Metode Jaccard. JITTER : Jurnal Ilmiah Teknologi dan Komputer, [S.l.], v. 3, n. 1, p. 786-793, jan. 2022. ISSN 2747-1233. Available at: <https://ojs.unud.ac.id/index.php/jitter/article/view/82747>. Date accessed: 10 dec. 2022.

[9] Putra, I. M. S., Putu Jhonarendra, & Ni Kadek Dwi Rusjayanthi. (2021). Deteksi Kesamaan Teks Jawaban pada Sistem Test Essay Online dengan Pendekatan Neural Network . Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 5(6), 1070 - 1082. https://doi.org/10.29207/resti.v5i6.3544