OPTIMASI ALGORITMA VECTOR SPACE MODEL DENGAN ALGORITMA K-NEAREST NEIGHBOUR PADA PENCARIAN JUDUL ARTIKEL JURNAL

Siti Fauziah[1]; Daning Nur Sulistyowati[2]; Taufik Asra[3]

[1]Program Studi Ilmu Komputer
STMIK Nusa Mandiri Jakarta
www.nusamandiri.ac.id
[1]sitifauziah478@gmail.com; 2dns9321@gmail.com

[3]Program Studi Rekayasa Perangkat Lunak
Universitas Bina Sarana Informatika
www.bsi.ac.id
taufik.tas@bsi.ac.id

Abstract— Articles is one part of the scientific work which was manifested in the form of writing and containing a lot of information that are requisite and suited therein to the exclusion of .Many small article day with allah is as a variety of sorts of the title and the methodology that was used , but does not make up for the possibility of a resemblance of the title of the article that is there is .This study aims to for determining the rate of a resemblance between an article of the american journal of public from the point of view of the title of the articles the american journal of public by the use of an algorithm of vector space the model and compare it with an algorithm k-nearest neghbour .The data used pt pgn promised to supply 10 the title of an article of the american journal of public keyword on information retrieval .Testing the data with of these keywords documents produced by the only by the magnitude of the resemblance of its on the highest a method of vsm it will be on a doc 5 , doc 7 , doc 8 and doc 4 .While for the program knn generate a level of the resemblance of its on range doc7 , doc10| doc8 , doc10| doc4 , d10| doc5 , doc10| doc3 , doc10. So that came to the conclusion that the occurrence of the addition of the criteria used to they obtain documents they do similaritas keyword after using an algorithm k-nearest neghbour.

***Keywords**: Journal Title Search, Vector Space Model (VSM), K-Nearest Neghbour (KNN)*

Abstrak— Artikel merupakan salah satu bentuk karya ilmiah yang dituangkan dalam bentuk tulisan dan mengandung banyak informasi yang berguna didalamnya. Banyak artikel yang ada dengan berbagai macam judul dan metode yang digunakan, namun tidak menutup kemungkinan adanya kemiripan dari judul artikel yang ada. Penelitian ini bertujuan untuk menentukan tingkat kemiripan antara artikel jurnal dilihat dari judul artikel jurnal dengan menggunakan *algoritma vector space model* dan membandingkannya dengan *algoritma k-nearest neghbour*. Data yang digunakan yaitu 10 judul artikel jurnal dengan kata kunci *Information Retrieval*. Pengujian data dengan kata kunci tersebut menghasilkan dokumen dengan tingkat kemiripan tertinggi pada metode VSM yaitu pada Dok 5, Dok 7, Dok 8 dan Dok 4. Sedangkan untuk KNN menghasilkan tingkat kemiripan pada range Doc7,Doc10 | Doc8,Doc10 | Doc4,D10 | Doc5,Doc10 | Doc3,Doc10. Sehingga menyimpulkan bahwa terjadinya penambahan kriteria dokumen yang similaritas dengan kata kunci setelah menggunakan *algoritma K-Nearest Neghbour*.

***Kata Kunci**: Pencarian Judul Jurnal, Vector Space Model (VSM), K-Nearest Neghbour (KNN)*

PENDAHULUAN

Artikel merupakan salah satu bentuk karya ilmiah yang dituangkan dalam bentuk tulisan. Dalam artikel jurnal dapat mengandung informasi-informasi yang berguna bagi para pembaca. Banyak penulis membuat artikel jurnal dengan judul yang beraneka ragam, namun tidak menutup kemungkinan adanya kemiripan dari judul artikel yang dibuat.

Kemiripan suatu artikel jurnal dapat dilihat hanya dari judul artikel tersebut. Tetapi tidak dapat menentukan tingkat kemiripan tertinggi dari masing-masing dokumen yang ditemui.

Metode penelitian yang digunakan yaitu mempelajari teori-teori literatur dan buku-buku yang berhubungan dengan objek kajian, melakukan kajian baik secara online maupun offline dan menganalisa data. Data yang dikumpulkan selanjutnya dilakukan analisa dan perbandingan secara manual dengan menggunakan dua algoritma.

Dalam penelitian yang dikembangkan oleh (Mas`udia Putri Elfa, Atmadja, Martono Dwi, Mustafa, 2017) suatu sistem temu kembali informasi judul tugas akhir dan perhitungan kemiripan dokumen menggunakan vector space model. Sistem secara otomatis akan melakukan indexing secara offline dan temu kembali (retrieval) secara real time.

Sedangkan dalam penelitian (Wisnu & Hetami, 2015) memanfaatkan information retrieval pada text mining untuk menemukan ide pokok dalam teks pada artikel berbahasa inggris untuk membantu pembaca untuk lebih mudah memahami isi artikel dan menghemat waktu yang dibutuhkan.

Tujuan dari dibuatnya penelitian ini yaitu untuk menentukan tingkat kemiripan antara artikel jurnal dilihat dari judul artikel jurnal dengan menggunakan *algoritma vector space* model dan membandingkannya dengan *algoritma k-nearest neghbour.*

BAHAN DAN METODE

A. *Information Retrieval* (Sistem Temu Kembali)

Information Retrieval merupakan sistem yang menerima *query* dari pengguna, kemudian dilakukan *ranking* terhadap dokumen berdasar kesesuian terhadap *query*. Hasil *ranking* yang diberikan pada pengguna merupakan dokumen yang menurut sistem memiliki relevansi terhadap *query*, tetapi tingkat relevansi itu sendiri merupakan hal yang subjektif tergantung dari pengguna yang dipengaruhi oleh berbagai macam faktor seperti topik, pewaktuan, sumber informasi maupun tujuan pengguna. Model sistem temu kembali menentukan detail sistem temu yaitu meliputi *representasi* dokumen maupun *query*, fungsi pencarian (*retrieval function*), dan notasi kesesuaian (*relevance notation*) dokumen terhadap *query*.

Menurut (Bunyamin & Negara, 2016) *Information Retrieval* terbagi dari beberapa bagian yang dijabarkan sebagai berikut:

1. *Text Operations*, meliputi pemilihan kata-kata dalam querymaupun dokumen (*term selection*) dalam proses transformasi dokumen atau query menjadi *term index* (indeks kata-kata).
2. *Query formulation*, memberi bobot pada indeks kata-kata *query*.
3. *Ranking*, mencari dokumen-dokumen yang relevan terhadap *query* dan mengurungkan dokumen tersebut berdasarkan kesesuaiannya dengan *query*.
4. *Indexing*, membangung basis data indeks dari koleksi dokumen dilakukan terlebih dahulu sebelum pencarian dokumen dilakukan.

B. Text Mining

Menurut (Amburika, Chrisnanto, & Uriawan, 2016) *text mining* merupakan salah satu bidang khusus dari data mining. *Text mining* dapat didefiniskan sebagai suatu proses menggali informasi dimana seorang *user* berinteraksi dengan sekumpulan dokumen menggunakan *tool* analisis yang merupakan komponen-komponen dalam data mining. Dalam *text mining* berbeda dengan dengan data mining dimana data mining yang digunakan adalah *structured* data sementara dalam *text mining* umumnya data yang ditemui adalah *semi-structured* atau *unstructured*. Sementara keduanya memiliki permasalah yang sama yaitu jumlah data yang besar, dimensi yang tinggi, dan data juga struktur yang terus berubah. Struktur teks yang kompleks dan tidak lengkap, arti yang tidak jelas dan tidak standar, dan bahasa yang berbeda ditambah terjemahan yang tidak akurat memberikan tantangan tambahan pada *text mining. Text mining* dalam prakeknya mencari pola-pola tertentu, mengasosiasikan satu bagian teks dengan lain berdasar aturan-aturan tertentu, kata-kata yang dapat mewakili sehingga dapat dilakukan analisa keterhubungan antar satu dengan lain, dalam kumpulan dokumen yang sangat banyak. Dokumen yang ada bisa bersifat statis, yaitu dokumen yang tidak akan di perbarui lagi ataupun dinamis yaitu dokumen yang akan selalu diperbarui dalam rentang waktu tertentu.

Tahapan *Text Mining*

a. Case Folding
Mengubah semua huruf dalam dokumen menjadi huruf kecil (*lowercase*). Dalam tahap ini juga karakter selain huruf dihilangkan.

b. Tokenizing
Memotong tiap kata dalam kalimat atau parsing dengan menggunakan spasi sebagai delimiter yang akan menghasilkan token berupa kata.

c. Filtering

Menyaring kata yang didapat dari proses tokenizing yang dianggap tidak penting atau tidak memiliki makna dalam proses text mining yang disebut *stoplist*. Tiap kata yang diperoleh dari *tokenizing* akan dicocokkan dalam kamus *stopword* di dalam *database*, jika kata tersebut cocok dengan salah satu kata dalam *stopword* maka kata tersebut akan dihilangkan, sementara yang tidak cocok akan dianggap cocok dan diproses ke tahap selanjutnya.

d. Stemming

Mengembalikan kata-kata yang diperoleh dari hasil *filtering* ke bentuk dasarnya, menghilangkan imbuhan awal (*prefix*) dan imbuhan akhir (*sufix*) sehingga di dapat kata dasar.

e. Tagging

Merubah kata dalam bentuk lampau (*past tense*) menjadi bentuk sekarang (*future tense*).

f. Analyzing

Keterhubungan antar kata dalam dokumen akan ditentukan dengan menghitung frekuensi term pada dokumen atau lebih sering dikenal dengan tahap pembobotan.

C. TF-IDF *(Term Frequency-Inverse Document Frequency)*

Basis pembobotan TF-IDF merupakan jenis pembobotan yang melibatkan pengukuran statistik untuk mengukur seberapa penting sebuah kata dalam kumpulan dokumen. Tingkat kepentingan meningkat ketika sebuah kata muncul beberapa kali dalam sebuah dokumen tetapi diimbangi dengan frekuensi kemunculan kata tersebut dalam kumpulan dokumen (Wisnu & Hetami, 2015). TF merupakan pembobotan yang sederhana dimana penting tidaknya sebuah kata diasumsikan sebanding dengan jumlah kemunculan kata tersebut dalam dokumen, sementara IDF merupakan pembobotan yang mengukur seberapa penting sebuah kata dalam dokumen apabila dilihat secara global pada seluruh dokumen (M.Isa & Abidin, 2013).

Perhitungan IDF menggunakan persamaan 1

$$IDF_{(t)} = \log(D \,/\, df_{(t)}) \quad \text{..... (1)}$$

Dimana:

$df_{(t)}$ = Jumlah dokumen yang mengandung kata ke-t dari kata kunci

D = Jumlah semua dokumen yang ada di dalam database

IDF = Rasio frekuensi dokumen pada kata ke-t dari kata kunci

Perhitungan TF-IDF menggunakan persamaan 2

$$TF - IDF_{(d,t)} = TF_{(d,t)} * IDF_{(t)} \text{..... (2)}$$

Dimana:

d = dokumenke-d

t = kata ke-t dari kata kunci

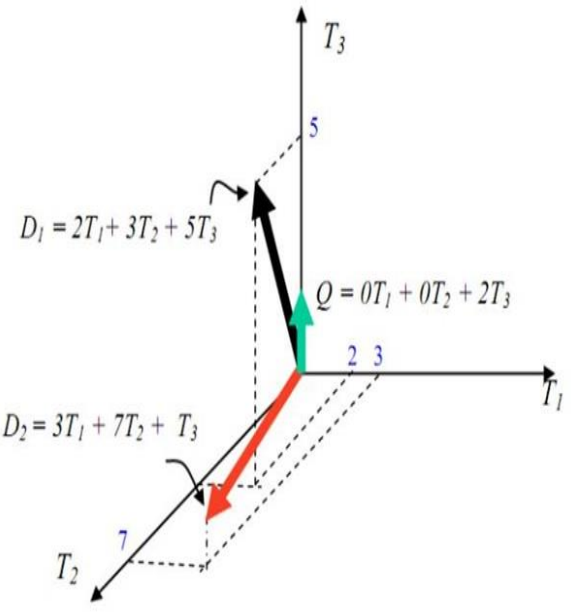
tf = frekuensi banyaknya kata ke-t dari kata kunci pada dpkumen ke-d

TF-IDF = bobot dokumen ke-d terhadapkata kunci ke-t

IDF = rasio frekunsi dokumen pada kata ke-t dari kata kunci

D. *Algoritma Vector Space Model*

Vector Space Model (VSM) merupakan model Information Retrieval yang mempresentasikan dokumen dan query sebagai vektor pada ruang multidimensi. Kesamaan suatu dokumen dengan query dapat diukur dengan vektor dokumen dan vektor query (Aziz, Saptono, & Suryajaya, 2015). Dalam metode *Vector Space Model* dihitung *weighted* dari setiap *term* yang terdapat dalam semua dokumen dan *query* dari *user. Term* adalah kata atau kumpulan kata yang merupakan ekspresi verbal dari suatu pengertian. Penentuan relevansi dokumen dengan *query* dipandang sebagai pengukuran kesamaan antara vektor dokumen dengan vektor *query*. Contoh representasi relevansi antara dokumen dan *query* dapat digambarkan pada Gambar 1. Q merupakan *query* pembanding, D1 dan D2 adalah dua dokumen yang akan dibandingkan, sedangkan T1, T2 dan T3 adalah tiga *term* pada dokumen tersebut.



Sumber: (Fauziah & Sulistyowati, 2018)
Gambar 1. *Representasi* dokumen dan *query* pada ruang vektor

Pada perhitungan VSM digunakan pembobotan TF-IDF dan perhitungan nilai *similarity* dengan menggunakan *Cosine Similarity*. Metode TF-IDF adalah cara untuk memberikan bobot hubungan suatu *term* terhadap dokumen. Metode ini menggabungkan dua konsep perhitungan bobot yaitu frekuensi kemunculan kata dalam suatu dokumen dan *inverse* dari frekuensi yang mengandung kata tersebut.

E. Algoritma K-Nearest Neghbour

Algoritma K-Nearest Neighbor (K-NN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. K-NN termasuk *algoritma supervised learning* dimana hasil dari *query instance* yang baru diklasifikan berdasarkan mayoritas dari kategori pada K-NN. Kelas yang paling banyak munculah yang akan menjadi hasil klasifikasi. Klasifikasi menggunakan *voting* terbanyak diantara klasifikasi dari k obyek. *Algoritma K-Nearest Neighbor* (K-NN) menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari *query instance* yang baru. Perhitungannya adalah dengan menjumlahkan semua nilai kemiripan yang tergolong dalam satu kategori kemudian membandingkan manakah yang lebih besar. Rumusnya adalah sebagai berikut:

$$p(x, c_m) = \sum_{l=1}^{m} SIM(X, d_j) \in c_m \quad \text{..... (3)}$$

Keterangan:

P(x,cm) : probabilitas dokumen X menjadi anggota kategori cm

sim(x,dj)∈cm : kemiripan antara dokumen X dengan dokumen latih dj yang merupakan anggota dari kategori cm

m : jumlah sim(x,dj) yang termasuk dalam kategori cm

HASIL DAN PEMBAHASAN

Pengujian yang dilakukan dengan mengunakan 10 judul artikel jurnal berbahasa Indonesia dengan *query Information Retrieval*. Data judul artikel yang akan diuji adalah sebagai berikut:

Tabel 1. Data Judul Artikel Jurnal

No	Judul Artikel Jurnal
1	Information Retrieval System Pada Pencarian File Dokumen Berbasis Teks Dengan Metode Vector Space Model Dan Algoritma ECS Stemmer
2	Klasifikasi Konten Berita Dengan Metode Text Mining
3	Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi
4	Perancangan Information Retrieval (IR) Untuk Pencarian Ide Pokok Teks Artikel Berbahasa Inggris Dengan Pembobotan Vector Space Model
5	Information Retrieval Tugas Akhir Dan Perhitungan Kemiripan Dokumen Mengacu Pada Abstrak Menggunakan Vector Space Model
6	Aplikasi Information Retrieval (IR) CATA Dengan Metode Generalized Vector Space Model
7	Rancang Bangun Information Retrieval System (IRS) Bahasa Jawa Ngoko pada Palintangan Penjebar Semangad dengan Metode Vector Space Model (VSM)
8	Implementasi Vector Space Model Dan Beberapa Notasi Metode Term Frequency Inverse Document Frequency (TF-IDF) Pada Sistem Temu Kembali Informasi
9	Efektifitas Seleksi Fitur Dalam Sistem Temu-Kembali Informasi
10	Klasifikasi Dokumen Tumbuhan Obat Menggunakan Metode Improved k-Nearest Neighbor

Sumber: (Fauziah & Sulistyowati, 2018)

Judul artikel jurnal tersebut akan mengalami proses text mining yang bertujuan untuk mempersiapkan text menjadi data yang akan mengalami pengolahan pada proses selanjutnya. Dan akan dilanjutkan pada proses perhitungan dimulai dari menghitung tfidf, menghitung jarak *query* dan menghitung similaritas dokumen.

Hasil Pengujian Klasifikasi dengan Algoritma Cosine Similarity Pada pengujian dalam tahap ini dilakukan 2 tahap pengujian, yaitu tahap training data dan tahap testing. Pada tahap training, data yang digunakan telah diketahui jenis kategorinya. Tahap training digunakan untuk melihat ketepatan klasifikasi dokumen dengan algoritma cosine similarity, pada tahap ini data

yang digunakan adalah sejumlah 10 dokumen. Dokumen yang dipilih merupakan dokumen dengan Query Information Retrieval. Data yang digunakan ditunjukkan pada Tabel I. Selanjutnya merupakan tahap testing, dimana data-data yang diolah belum diketahui kategorinya dan akan mendapatkan kategori secara otomatis dari sistem. Dalam penelitian ini, kategori yang dipakai adalah sejumlah 5 kategori dengan pemakaian kata kunci sedemikian rupa. Tabel II merupakan data hasil tahap testing.

Tabel 2. Menghitung Similaritas

No	Nama Dokumen	Keterangan
1	D7,D10	Valid
2	D8,D10	Valid
3	D4,D10	Valid
4	D5,D10	Valid
5	D3,D10	Valid
6	D1,D10	Tidak Valid
7	D2,D10	Tidak Valid
8	D6,D10	Tidak Valid
9	D9,D10	Tidak Valid

Sumber: (Fauziah & Sulistyowati, 2018)

Setelah dilakukan proses perhitungan tersebut maka akan diperoleh hasil seperti terlihat pada tabel berikut:

Tabel 3. Hasil *Similaritas* Dokumen Dengan Metode VSM

Dokumen	Nilai	Rank
D5	0,88	1
D7	0,22	2
D8	0,20	3
D4	0,19	4
D3	0,18	5
D1	0,17	6
D10	0,17	6
D2	0,15	7
D6	0,14	8
D9	0,13	9

Sumber: (Fauziah & Sulistyowati, 2018)

Sementara dengan metode KNN diperoleh hasil sebagai berikut dengan nilai K=4:

Tabel 4. Hasil Perhitungan Dengan Metode KNN

No	Nama Dokumen	Distance	Rangking
1	D7,D10	3,69	1
2	D8,D10	3,69	1
3	D4,D10	3,25	2
4	D5,D10	3,18	3
5	D3,D10	3,16	4
6	D1,D10	3,13	5
7	D2,D10	3	6
8	D6,D10	2,8	7
9	D9,D10	2,71	8

Sumber: (Fauziah & Sulistyowati, 2018)

KESIMPULAN

Berdasarkan hasil dari analisa dan pengolahan data pada pembahasan diatas, dapat ditarik kesimpulan sebagai berikut:
Dari hasil pengujian dengan kata kunci *Information Retrieval* maka akan menghasilkan dokumen dengan tingkat kemiripan tertinggi pada *Algoritma Vector Space Model* yaitu pada pada Dok 5, Dok 7, Dok 8 dan Dok 4. Sedangkan dengan *Algoritma K-Nearest Neghbour* menghasilkan tingkat kemiripan pada *range* Doc7,Doc10 | Doc8,Doc10 | Doc4,D10 | Doc5,Doc10 | Doc3,Doc10. Setelah dilakukan perhitungan dengan *Algoritma Vector Space Model* dan *Algoritma K-Nearest Neghbour* terjadi penambahan kriteria dokumen yang similaritas dengan kata kunci.

REFERENSI

Amburika, B., Chrisnanto, Y. H., & Uriawan, W. (2016). TEKNIK VECTOR SPACE MODEL (VSM) DALAM PENENTUAN PENANGANAN DAMPAK GAME ONLINE PADA ANAK. In *Prosiding SNST ke-7 Tahun 2016* (pp. 73–78).

Aziz, A., Saptono, R., & Suryajaya, K. P. (2015). Implementasi Vector Space Model dalam Pembangkitan Frequently Asked Questions Otomatis dan Solusi yang Relevan untuk Keluhan Pelanggan. *Scientific Journal of Informatics*, *Vol. 2*, *No*, 111–122.

Bunyamin, H., & Negara, C. P. (2016). Aplikasi Information Retrieval (IR) CATA Dengan Metode Generalized Vector Space Model. *Jurnal Informatika*, *4*.

Fauziah, S., & Sulistyowati, D. N. (2018). *Laporan Akhir Penelitian Mandiri - Optimasi Algoritma Vector Space Model Dengan Algoritma K-Nearest Neighbour Pada Pencarian Judul Artikel Jurnal*. Jakarta.

M.Isa, T., & Abidin, T. F. (2013). Mengukur Tingkat

Kesamaan Paragraf menggunakan Vector Space Model untuk Mendeteksi Plagiarisme. In *Seminar Nasional dan Expo Teknik Elektro*.

Mas`udia Putri Elfa, Atmadja, Martono Dwi, Mustafa, L. D. (2017). Information Retrieval Tugas Akhir Dan Perhitungan Kemiripan Dokumen Mengacu Pada Abstrak Menggunakan Vector Space Model. *Jurnal Teknik Mesin, Eektro Dan Ilmu Komputer*, *8*.

Wisnu, D., & Hetami, A. (2015). Perancangan Information Retrieval (IR) Untuk Pencarian Ide Pokok Teks Artikel Berbahasa Inggris Dengan Pembobotan Vector Space Model. *Jurnal Ilmiah Teknologi Dan Informasi ASIA*, *9*.