PENGEMBANGAN *VECTOR SPACE MODEL* PADA PENGUKURAN KEMIRIPAN PUBLIKASI ILMIAH

Stephanie Betha Rossi Hersianie
Program Studi Teknik Komputer, Fakultas Teknik, Universitas Wiralodra
email : ntephbetha@gmail.com

Abstrak

Publikasi ilmiah dapat mengandung lebih dari satu topik atau kategori bidang penelitian. Identifikasi topik atau bidang penelitian dapat dilihat hanya dari membaca judul publikasi ilmiah tersebut. Namun, judul publikasi ilmiah tidak dapat digunakan untuk menentukan kemiripannya dengan kategori bidang penelitian tertentu karena judul publikasi ilmiah belum tentu dapat mencerminkan bidang penelitiannya. Hal ini membuat pencarian judul publikasi ilmiah yang dilakukan oleh penulis jurnal menjadi kurang efektif. Kemiripan publikasi ilmiah dengan kategori bidang penelitiannya dapat ditentukan menggunakan *Vector Space Model*. Permasalahan pertama yang terjadi adalah skema pembobotan TFIDF pada *Vector Space Model* tidak dapat diimplementasikan pada penelitian ini. Penyebabnya adalah skema tersebut belum dapat mewakili kategori bidang penelitian. Selain itu, matriks pembobotan TFIDF juga memerlukan penyesuaian kolom untuk memproses dataset yang berjumlah besar. Permasalahan kedua yaitu pengukuran kemiripan dokumen antara *query* dengan panjang vektor dokumen. Panjang vektor dokumen pada penelitian sebelumnya diperoleh dari jumlah kata yang ada pada suatu dokumen. Sedangkan pada penelitian ini dibutuhkan pengukuran kemiripan dokumen yang berupa judul publikasi ilmiah dengan kategori bidangnya. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan *Vector Space Model* dalam mengukur kemiripan judul publikasi ilmiah dengan kategori bidang penelitiannya. Penelitian ini menghasilkan nila rata-rata recall sebesar 89,7 % dan presisi sebesar 90%.

Keyword : *Vector Space Model,Pembobotan TFIDF, Publikasi Ilmiah*

PENDAHULUAN

Publikasi ilmiah merupakan hasil dari proses penelitian dan pemikiran dari suatu gagasan dalam penyelesaian suatu masalah [1]. Publikasi ilmiah dapat mengandung lebih dari satu topik atau kategori bidang penelitian [2]. Identifikasi topik atau bidang penelitian dapat dilihat hanya dari membaca judul publikasi ilmiah tersebut. Namun, judul publikasi ilmiah tidak dapat digunakan untuk menentukan kemiripannya dengan kategori bidang penelitian tertentu secara langsung [3]. Judul publikasi ilmiah belum tentu dapat mencerminkan bidang penelitiannya. Hal ini membuat pencarian judul publikasi ilmiah yang dilakukan oleh penulis jurnal menjadi kurang efektif, karena penulis jurnal harus membaca isi publikasi ilmiah secara keseluruhan untuk menentukan kategori bidang penelitian pada publikasi tersebut. Oleh karena itu, diperlukan metode untuk mengukur kemiripan antara judul publikasi ilmiah dengan kategori bidang penelitiannya. Kemiripan publikasi ilmiah dengan kategori bidang penelitiannya dapat ditentukan menggunakan *Vector Space Model*.

Hal ini disebabkan oleh *Vector Space Model* terdiri atas proses pengindeksan, pembobotan dokumen, pemberian rangking antar dokumen untuk mengukur tingkat kesamaan. Pembobotan dokumen dilakukan menggunakan pembobotan TFIDF. Perangkingan dilakukan dengan mengukur kemiripan antara judul publikasi ilmiah (vektor *query*) dengan kategori bidang penelitian (panjang vektor dokumen). Pengukuran kemiripan ini menggunakan algoritma *cosine similarity*. Dedi dkk [8] telah menerapkan algoritma *cosine similarity* dan pembobotan TFIDF pada sistem arsip dokumen. Namun penelitian ini hanya menggunakan jumlah data yang kecil. Sedangkan untuk mengukur kemiripan judul publikasi ilmiah dengan kategori bidangnya, maka dibutuhkan jumlah data yang besar. Jumlah data yang besar ini digunakan untuk menghasilkan panjang vektor dokumen yang dapat mewakili setiap kategori bidang penelitian. Implementasi jumlah data yang berbeda dapat menimbulkan beberapa permasalahan pada *Vector Space Model*.

Permasalahan pertama yang terjadi adalah skema umum pembobotan TFIDF pada *Vector Space Model* biasanya disajikan dengan kolom yang terdiri atas *term* atau kata pada setiap dokumen (kalimat), kolom jumlah frekuensi kata pada dokumen kesatu, kedua dan seterusnya [7]. Matriks pembobotan tersebut, khususnya pada kolom dokumen, dianggap belum dapat mewakili kategori bidang penelitian. Untuk menentukan kategori bidang penelitian, dibutuhkan pembangunan model yang dapat mewakili setiap kategori bidang penelitian. Pembangunan model dilakukan dengan menghitung panjang vektor kategori setiap bidang penelitian. Input dari tahapan ini adalah dokumen latih [9]. Permasalahan kedua yaitu pengkuran kemiripan dokumen antara *query* dengan panjang vektor dokumen dengan *cosine similarity*. Panjang vektor dokumen pada penelitian [8] diperoleh dari jumlah kata yang ada pada suatu dokumen. Sedangkan pada penelitian ini dibutuhkan pengukuran kemiripan dokumen yang berupa judul publikasi ilmiah dengan kategori bidangnya. Judul publikasi ilmiah merupakan *query* dan kategori bidang diperoleh dari panjang vektor dokumen. Panjang vektor dokumen ini tidak hanya diperoleh dari jumlah kata yang ada pada suatu dokumen, melainkan jumlah kata yang ada pada kumpulan dokumen di suatu kategori. Oleh karena itu, pengukuran kemiripan dokumen pada penelitian ini diperoleh dari perhitungan *cosine similarity* antara panjang vektor kategori dengan *query* dokumen. *Query* dokumen merupakan dokumen uji yang berupa judul publikasi ilmiah [9]. Sedangkan, kategori bidang diperoleh dari model yang telah dibangun pada proses pembobotan.

Berdasarkan permasalahan di atas, maka dibutuhkan pengembangan *Vector Space Model*. Rumusan masalah dari penelitian ini yaitu, bagaimana hasil pengembangan *Vector Space Model* yang digunakan dalam pengukuran kemiripan judul publikasi ilmiah. Tujuan dari penelitian ini adalah mengetahui ketepatan pengembangan *Vector Space Model* dalam pengukuran kemiripan judul publikasi ilmiah. Manfaat dari penelitian ini adalah diharapkan menghasilkan model pengukuran kemiripan publikasi ilmiah. Batasan masalah dari penelitian ini adalah

kategori bidang penelitian yang digunakan mengikuti standar klasifikasi *Field of Research* (FoR) di bidang *Information and Computing Science* berjumlah 7 kategori bidang. Atribut dokumen publikasi ilmiah yang digunakan dalam pengukuran kemiripan publikasi ilmiah adalah atrribut judul.

Penelitian terdahulu pernah mengangkat topik serupa dengan penelitian ini sehingga dapat dijadikan acuan dalam pelaksanaan penelitian ini. Beberapa penelitian terdahulu telah menerapkan *cosines similarity* dalam pengukuran kemiripan dokumen, serta menerapkan *Vector Space Model* dalam perangkingan suatu dokumen. Penelitian pertama yaitu penelitian Dedi dkk yang menerapkan algoritma *cosine similarity* pada sistem arsip dokumen di Universitas Islam Sultan Agung. Hasil dari penelitian tersebut adalah algoritma *cosine similarity* mampu menemukan dokumen arsip dengan tingkat kemiripan yang tinggi [8]. Hal ini ditunjukkan dengan menggunakan pengukuran kinerja algoritma *cosine similarity* menunjukkan angka *precision* 88,8% dan *recall* 76.1%. Namun, penelitian tersebut hanya menggunakan jumlah dokumen yang tidak terlalu banyak sehingga dibutuhkan pengembangan terhadap *Vector Space Model* yang digunakan dalam penelitian ini.

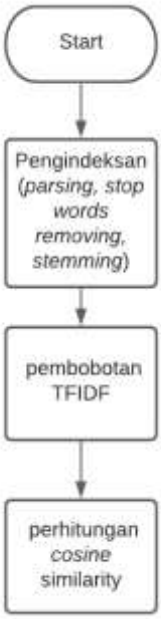
Penelitian selanjutnya yaitu penelitian Sekarwati dkk yang menerapkan perhitungan *cosine similarity* dalam pengukuran kemiripan dokumen. Hasilnya adalah presentase kemiripan *cosine similarity* hampir mendekati 100% [4]. Hal ini membuktikan bahwa *cosine similarity* cukup akurat digunakan dalam pengukuran kemiripan dokumen. Selain itu, Yisti, V dan Retno, M juga melakukan modifikasi algoritma *cosine similarity* untuk deteksi kemiripan dokumen skripsi mahasiswa. *Cosine Similarity* dikombinasikan dengan *Levenshtein Distance* yang menghasilkan suatu aplikasi berbasis website [5]. Kombinasi ini digunakan agar dapat mendeteksi *plagiarisme* dengan lebih akurat. Penelitian Eva dkk juga menerapkan *Vector Space Model* serta membandingkannya metode *Winnowing* dengan dalam pengukuran kemiripan dokumen. Hasilnya adalah *Vector Space Model* mempunyai langkah-langkah yang lebih praktis dibandingkan metode *Winnowing* [6]. Walaupun memiliki langkah sederhana, tetapi *Vector Space Model* tetap memiliki keakuratan yang tinggi.

Penelitian selanjutnya adalah penelitian Aziz, Abdul dan Bayu, Indra yang mengimplementasikan *Vector Space Model* dengan pembobotan TF-IDF dalam pencarian dokumen. Konsep yang digunakan adalah mengetahui kedekatan dua buah vektor, yaitu dengan cara menghitung besarnya sudut di antara kedua vektor tersebut. Hasil rangking diperoleh dari pengurutan sudut terbesar sampai terkecil [9]. Penelitian tersebut juga menerapkan konsep data latih dan data uji dalam perhitungan akurasinya.

Betha, Stephanie juga menerapkan *Vector Space Model* dalam penentuan *multiple membership* dokumen. Setiap dokumen mengandung beberapa kategori bidang. *Vector Space Model* membantu dalam meranking bidang tersebut, semakin tinggi presentase suatu bidang, maka semakin relevan suatu dokumen terhadap kategori bidang tersebut [2]. Berdasarkan penelitian-penelitian yang telah dipaparkan di atas, penelitian ini berfokus mengembangkan metode *Vector Space Model* pada pengukuran kemiripan dokumen publikasi ilmiah. Pengembangan *Vector Space Model* dilakukan pada tahap skema pembobotan TFIDF dan pengukuran kemiripan dokumen dengan menentukan panjang vektor kategori di setiap kategori dokumen. Panjang vektor kategori dilakukan melalui pembangunan model. Input dari pembangunan model ini adalah judul dokumen publikasi ilmiah berbahasa Inggris berjumlah 3299. Dokumen tersebut dibagi menjadi dua yaitu dokumen latih dan dokumen uji [9]. Dokumen latih telah memiliki label kategori. Data ini diperoleh dari jurnal pada *Computer Science Bibliography* (DBLP) dan KK RPL STEI ITB yang telah diberi label secara manual.

METODE PENELITIAN

Pada bagian ini dijelaskan tentang metode yang digunakan untuk pengukuran kemiripan judul publikasi ilmiah menggunakan pengembangan *Vector Space Model*. Pada penelitian ini metode yang dibahas terdiri dari analisis sistem yang terdiri dari analisis masalah dan solusi. *Vector Space Model* memiliki beberapa tahapan yaitu pengindeksan, pembobotan dan perhitungan *cosine similarity* untuk mengukur kemiripan dokumen. Proses pengindeksan terdiri dari *parsing*, *stop words removing* dan *stemming*, sedangkan proses perangkingan terdiri dari pengukuran *similarity* dokumen menggunakan *cosine similarity* [11]. Gambar 1. menjelaskan tentang proses *Vector Space Model*.



Gambar 1. Alur Tahapan *Vector Space Model* Secara Umum

Pengindeksan terdiri dari beberapa proses yaitu proses yang pertama adalah *parsing*. *Parsing* merupakan proses mengambil *term-term* dari dokumen dengan cara memotong kata. Proses kedua adalah *stopwords removing*. Proses ini dilakukan dengan menghapus kata-kata yang tidak memiliki arti. Setelah itu, proses ketiga adalah *stemming*. *Stemming* merupakan proses pengembalian kata ke bentuk dasarnya. Proses pengindeksan ini akan menghasilkan daftar kata kunci. Selanjutnya, tahapan kedua adalah pembobotan TFIDF yaitu proses pembobotan kata. Daftar kata kunci yang telah mengalami pengindeksan dibobotkan sesuai dengan tingkat kepentingannya. Metode TFIDF 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 [11]. Rumus pembobotan TF-IDF tersebut [2] [10]:

$$Term\ Weight : w_i = tf_i * \log\left(\frac{D}{df_i}\right) \quad (1)$$

Tahapan ketiga adalah perangkingan. Perangkingan ini terdiri dari dua proses yaitu perhitungan vektor dokumen dan vektor *query*, serta perhitungan kemiripan antara vektor dokumen dengan vektor *query* menggunakan *cosine similarity*. Berikut ini adalah perhitungan vektor dokumen $|D|$ dan vektor *query* $|Q|$:

$$|D| \quad : \sqrt{\quad} \quad \sum_{i=1}^{n} Di^2 \quad (2)$$

$$|Q| \quad : \sqrt{\quad} \quad \sum_{i=1}^{n} Qi^2 \quad (3)$$

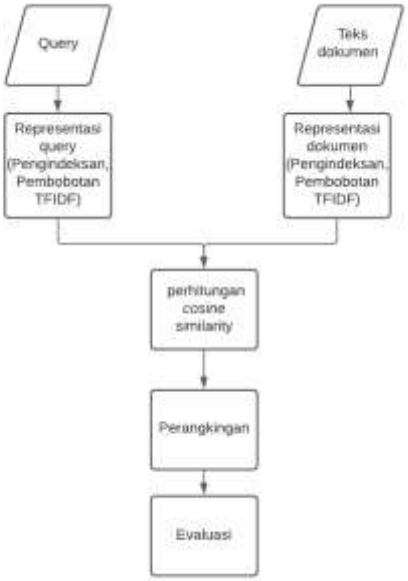
Perhitungan vektor dokumen pada rumus di atas adalah akar dari jumlah kuadrat kata yang ada di suatu dokumen. Sedangkan perhitungan vektor *query* di atas diperoleh dari akar dari jumlah kuadrat kata yang ada pada suatu *query*. Setelah itu, perhitungan *similarity* antara vektor dokumen dan vektor *query* adalah sebagai berikut :

$$Cos\ (Q, D) \quad : \frac{Q * D}{|Q| * |D|} \quad (4)$$

Perhitungan *cosine similarity* antara dokumen dan *query* akan menghasilkan nilai *cosinus*. Semakin besar nilai cosinus maka dianggap semakin dekat atau mirip suatu *query* terhadap dokumen. Pengurutan nilai tersebut secara *descending* inilah yang disebut sebagai proses perangkingan. Bobot yang besar artinya *query* tersebut memiliki kedekatan yang erat dengan dokumen [15].

Gambar 2. menjelaskan tahapan *Vector Space Model* secara detail. Tahapan tersebut terdiri dari teks dokumen mengalami pengindeksan dan pembobotan TFIDF, untuk menghasilkan representasi dokumen, selanjutnya, dilakukan pengajuan *query* terhadap sistem, *query* tersebut direpresentasikan melalui pengindeksan dan pembobotan TFIDF. Setelah itu, proses perhitungan *cosine similarity* antara *query* yang diajukan terhadap dokumen. Hasil perhitungan ini adalah nilai cosinus dari *query* terhadap dokumen. Dari nilai cosinus inilah, dapat diperoleh kedekatan *query* terhadap dokumen melalui perankingan. Kemudian, banyaknya *query* yang memiliki kedekatan dengan dokumen dihitung melalui tahapan evaluasi. Evaluasi atau pengujian terdiri dari perhitungan *recall* dan presisi. *Recall* adalah jumlah dokumen relevan yang terambil dibagi dengan jumlah dokumen relevan dalam database. Sedangkan, presisi adalah jumlah dokumen relevan yang terambil dibagi jumlah dokumen yang terambil dalam pencarian [13]. *Recall* juga dapat didefinisikan sebagai kemampuan sistem dalam menemukan hasil sesuai dengan *query* yang diajukan. Presisi merupakan ketepatan sistem untuk tidak menampilkan dokumen yang tidak sesuai dengan *query* yang diajukan [14]. Tahapan *Vector Space Model* secara detail digambarkan sebagai berikut

[12] :



Gambar 2. Tahapan *Vector Space Model* Secara Detail

Vector Space Model akan digunakan untuk mengukur kemiripan judul publikasi ilmiah terhadap kategori bidang penelitiannya. Kategori bidang penelitian diperoleh dari kumpulan dokumen latih yang memiliki label kategori bidang tertentu. Adanya jumlah dokumen yang banyak ini, maka timbul permasalahan pada saat menerapkan *Vector Space Model*. Permasalahan pertama yang terjadi adalah skema pembobotan kata pada TFIDF belum dapat mewakili kategori bidang penelitian. Identifikasi kategori bidang ini dibutuhkan untuk menentukan jenis kategori bidang penelitian pada suatu judul publikasi ilmiah. Karena pada pembobotan TFIDF hanya dapat menghitung jumlah kata pada suatu dokumen. Sedangkan yang dibutuhkan adalah perhitungan jumlah kata pada kumpulan dokumen yang memiliki label kategori tertentu.

Permasalahan kedua yang terjadi adalah pengkuran kemiripan dokumen. Pengukuran ini dilakukan dengan menghitung nilai *cosine similarity* antara *query* dengan panjang vektor dokumen. Perhitungan panjang vektor dokumen pada pengukuran kemiripan dokumen hanya jumlah akar kuadrat dari seluruh kata pada suatu dokumen. Namun, penelitian ini membutuhkan panjang vektor dokumen berasal dari kumpulan dokumen yang memiliki label kategori tertentu. Oleh karena itu, pada penelitian ini istilah panjang vektor dokumen diganti dengan panjang vektor kategori.
Tahapan *Vector Space Model* pada penelitian ini dilakukan secara garis besar sebagai berikut :

Gambar 3. Alur Tahapan *Vector Space Model* Pada Penelitian

Gambar 3 menjelaskan tentang alur tahapan *Vector Space Model* pada penelitian ini. Tahapan *Vector Space Model* yang digunakan pada penelitian ini mengalami beberapa pengembangan pada bagian pembobotan TFIDF dan pengukuran kemiripan dokumen pada perhitungan *cosine similarity* nya. Tahapan pertama yang dilakukan adalah tahap pengindeksan. Tahapan ini terdiri dari beberapa proses, yaitu, proses parsing, *stopwords removing* dan *stemming*. Proses parsing dilakukan dengan memotong judul publikasi menjadi daftar kata. Judul publikasi ilmiah ini berbentuk kata dalam berbahasa Inggris. Proses selanjutnya adalah proses penghapusan kata yang tidak memiliki arti pada susunan judul, misalnya, kata *in, and, for*. Setelah itu, kata tersebut mengalami *stemming*, misalnya, *removing* menjadi *remove*. *Controlling* menjadi *control* dan lainnya. Proses pengindeksan ini menghasilkan daftar kata.

Tahapan kedua adalah modifikasi proses pembobotan TFIDF. Setelah mengalami pengindeksan, daftar kata juga akan akan mengalami pembobotan. Tabel 1 pembobotan hanya menghitung banyaknya *term* i yang muncul pada sebuah dokumen, sedangkan penelitian ini membutuhkan

bobot suatu *term* pada suatu kategori bidang. Oleh karena itu, Tabel 1 akan mengalami modifikasi skema pembobotan menjadi Tabel 2 Pembobotan. Modifikasi pembobotan TFIDF ini bertujuan untuk menghitung banyaknya *term* yang muncul pada kumpulan dokumen di suatu kategori bidang. Modifikasi ini dilakukan pada skema tabel pembobotan sebagai berikut :

Tabel 2. Modifikasi Pembobotan TFIDF

Daftar Kata Kunci	tfC_i						weight : $tfC * icfi$		
	C_1	C_2	C_3	cfi	C/cfi	$icfi$	C_1	C_2	C_3
Kata ke-1									
Kata ke-2									
Kata ke-3									

$$Term\ Weight : w_i = tfC_i * \log\left(\frac{C}{cfi}\right) \quad (4)$$

Tabel 2 modifikasi pembobotan TFIDF menjelaskan bahwa bobot (w) dihasilkan dari perkalian banyaknya *term* i yang muncul pada suatu kategori dikalikan dengan nilai (icf). Nilai *icfi* berasal dari log $(C/cf_{(i)})$, dimana *C* adalah jumlah kategori dan *cfi* jumlah kategori yang memiliki kata tertentu. *C* merupakan kumpulan dari banyak dokumen (judul publikasi ilmiah) yang telah memiliki label kategori tertentu.

Tahapan ketiga adalah proses perangkingan. Perangkingan ini terdiri dari perhitungan vektor dokumen dan vektor *query*, serta perhitungan *cosine similarity* antara vektor dokumen dengan vektor *query*. Karena terjadi modifikasi pada skema pembobotan TFIDF, maka perhitungan vektor dokumen berubah menjadi perhitungan vektor kategori, sebagai berikut :

$$|C| \quad : \sqrt{\quad} \sum_{i=1}^{n} Ci^2 \quad (5)$$

Nilai vektor kategori adalah akar dari jumlah kuadrat kata atau *term* pada suatu kategori. Selanjutnya, perhitungan vektor *query* adalah sebagai berikut

$$|Q| \quad : \sqrt{\quad} \sum_{i=1}^{n} Qi^2 \quad (6)$$

Vektor *query* adalah akar dari jumlah kuadrat kata atau *term* yang ada pada suatu *query*. *Query* yang dimaksud dalam penelitian ini adalah judul publikasi ilmiah yang belum memiliki label kategori.

Setelah nilai vektor kategori dan vektor *query* diperoleh, maka tahapan selanjutnya adalah perhitungan nilai *cosine similarity*, sebagai berikut :

$$Cos\ (Q, C) \quad : \frac{Q * C}{|Q| * |C|} \quad (7)$$

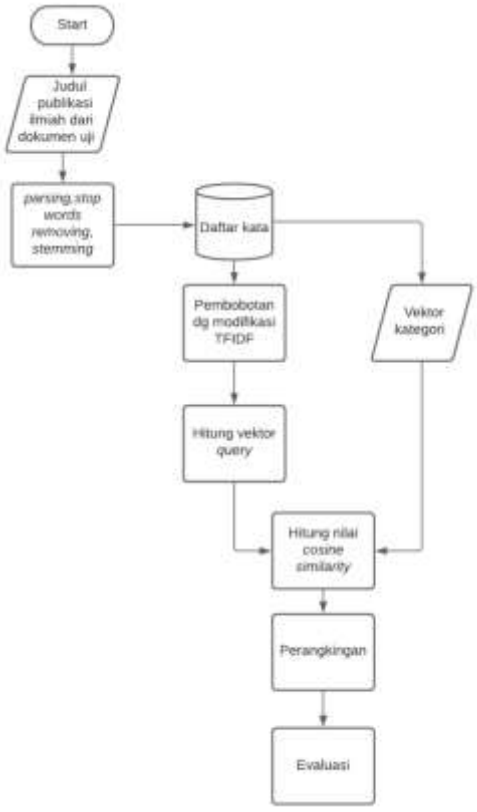
Nilai *cosine similarity* yang dihasilkan merupakan nilai *cosine similarity* (kedekatan vektor) antara setiap kategori dengan *query* yang diajukan. Semakin tinggi nilai *cosine similarity* pada suatu kategori artinya semakin dekat vektornya terhadap vektor *query*. Hal ini dapat diartikan bahwa *query* yang diajukan termasuk dalam kategori tersebut.

Proses implementasi pengembangan *Vector Space Model* pada penelitian ini terbagi menjadi dua bagian yaitu pembangunan model dan pengujian model. Proses pembangunan model menggunakan dokumen latih sedangkan proses pengujian model menggunakan dokumen uji. Proses pembangunan model merupakan proses pengindeksan, pembobotan TFIDF, perhitungan *cosine similiarity* pada judul publikasi ilmiah yang berasal dari dokumen latih. Proses ini diawali dengan pengindeksan pada dokumen latih. Setelah itu, hasil dari proses pengindeksan adalah daftar kata yang tersimpan dalam database. Daftar kata tersebut mengalami pembobotan TFIDF. Setelah dibobotkan, daftar kata tersebut akan dihitung jumlah vektor kategorinya sesuai dengan masing-masing kategori bidang.

Setelah proses pembangunan model selesai. Proses selanjutnya adalah proses pengujian model. Proses pengujian model yang telah dibangun dilakukan dengan proses pengajuan *query* berupa kata judul yang berasal dari dokumen uji. Gambar 4. Menjelaskan tentang proses pengujian model, yang terdiri dari proses pengindeksan *query*, pembobotan TFIDF pada *query*, proses perhitungan vektor *query*, proses perhitungan *cosine similarity*, proses perangkingan dan evaluasi.

Gambar 5. menjelaskan tentang detail proses pengujian model. Pengujian model diawali dengan pengindeksan judul publikasi ilmiah dari dokumen uji. Kemudian, hasil pengindeksan di simpan dalam database. Selamjutnya, daftar kata dalam database mengalami pembobotan TFIDF dan perhitungan vektor *query*. Setelah itu, perhitungan *cosine similarity* dilakukan antara vektor *query* dan vektor kategori. Vektor kategori ini diperoleh dari proses pembangunann model yang telah tersimpan di database. Selanjutnya, akan muncul beberapa nilai *similarity* dari masing-masing kategori bidang. Nilai ini telah diurutkan dari nilai tebesar sampai nilai terkecil (perangkingan). Evaluasi dilakukan dengan

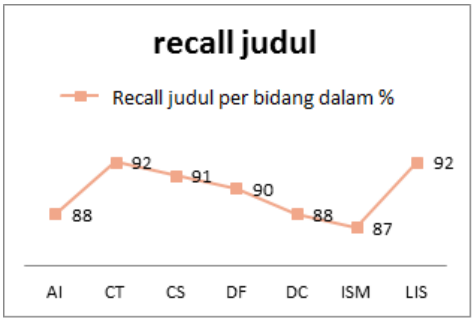
menguji coba semua dokumen uji untuk diproses pada tahapan pengujian model. Evaluasi ini menghasilkan *recall* dan presisi. Diagram alur proses pengujian model dijelaskan pada Gambar 5. di bawah ini :



Gambar 5. Hasil Pengujian *Recall* pada Judul

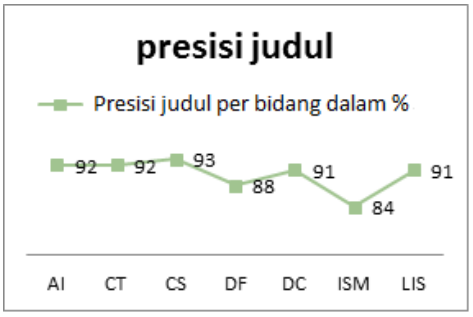
HASIL DAN PEMBAHASAN

Berikut ini adalah hasil pengujian *recall* dan presisi pada daftar kata judul publikasi ilmiah yang telah mengalami proses dengan modifikasi *Vector Space Model* :



Gambar 10. Hasil Pengujian *Recall* pada Judul

Gambar 10. menunjukkan tentang hasil *recall* pada judul publikasi ilmiah. Proses pengukuran kemiripan dokumen pada kategori bidang menggunakan pengembangan *Vector Space Model* menghasilkan nilai rata-rata sebesar 89,7%. Hal ini menunjukkan bahwa pengembangan *Vector Space Model* memiliki akurasi yang tinggi.



Gambar 11. Hasil Pengujian Presisi pada Judul

Gambar 11. menunjukkan tentang hasil presisi pada judul publikasi ilmiah. Proses pengukuran kemiripan dokumen pada kategori bidang menggunakan pengembangan *Vector Space Model* menghasilkan nilai rata-rata sebesar 90%. Kategori ketiga yaitu CS menghasilkan nilai presisi tertinggi dibandingkan kategori lain sebesar 93%. Hal ini disebakan oleh *query* yang diajukan banyak yang memiliki kemiripan dengan bidang CS.

KESIMPULAN

Pengembangan *Vector Space Model* dapat menghasilkan rata-rata nilai *recall* dan presisi yang tinggi, yaitu 89,7% dan 90% artinya, modifikasi *Vector Space Model* memiliki kinerja yang tinggi. Penelitian selanjutnya diharapkan dapat menggunakan atribut publikasi ilmiah lainnya, misalnya, abstrak atau isi publikasi ilmiah, sehingga lebih banyak memiliki variasi kata. Variasi kata yang lebih banyak dapat mempengaruhi tingkat akurasi pengembangan *Vector Space Model*.

DAFTAR PUSTAKA

[1] Sejati, FB dkk,"Deteksi Plagiarisme Karya Ilmiah dengan Pemanfaatan Daftar Pustaka Dalam Pencarian Kemiripan Tema Menggunakan Cosine Similarity", Jurnal Komtika. ,Vol.2 No.2, hal.85-94, 2019.

[2] Betha, Stephanie, "Penentuan Multimembership Dokumen", Majalah Ilmiah Unikom, Vol.15 No.2, hal.211-220, 2017.

[3] Fauziah, Siti dkk, Optimasi Algoritma *Vector Space Model* Dengan Algoritma K- Nearest Neighbour Pada Pencarian Judul Artikel Ilmiah, Jurnal PILAR Nusa Mandiri, Vol.15 No.1, hal.21-26.

[4] Sekarwati dkk, "Pengukuran Kemiripan Dokumen Menggunakan *Tools Gensim* ", Prosiding SNST ke-6, 2015.

[5] Yisti, V dan Retno, M, "Deteksi Kemiripan Dokumen Publikasi Skripsi Mahasiswa Menggunakan Algoritma Modifikasi *Cosine Similarity ", JIEET, Volume 03 Nomor 02, 2019.*

[6] Eva dkk, "A Comparison of Vector Space Model Method and Winnowing Algorithm to Measure the Similarity of Documents", The 5 th International Conference on Information Technology and Bussiness, 2019.

[7] Sharma, A, "Information System using Word2vec based Vactor Space Model ", Internet: https://www.analyticsvidhya.com/blog/2020/08/information-retrieval-using-word2vec-based-vector-space-model/, 2020 [Oct 15, 2021].

[8] Dedi dkk, "Implementasi Algoritma *Cosine Similarity* pada Sistem Arsip Dokumen di Universitas Islam Sultan Agung", TRANSFORMTIKA, Vol.17 No.2, pp. 124 – 132, 2020.

[9] Aziz, Abdul dan Bayu, Indra, "Implementasi *Vector Space Model* dalam Pencarian Dokumen", Prosiding Seminar Nasional Matematika dan Pendidikan Matematika, 2013.

[10] Amburika dkk, "Teknik *Vector Space Model* Dalam Penentuan Penanganan Dampak *Game Online* pada Anak", Prosiding Prosiding SNST ke-7, 2016.

[11] Aditya, Christian dan Nastiti, Vinna," Sistem Temu Kembali Buku Hadist Menggunakan Pembobotan *Term Frequency Inverse Document Frequency* dan *Cosine Similarity*", Seminar Nasional Teknologi dan Rekayasa (SENTRA), eISSN (Online) 2527-6050, 2019.

[12] Anna dan Hendini,Ana, "Implementasi *Vector Space Model* Pada Sistem Pencarian Karoke", Jurnal Evolusi, Vol. 6 No.1, 2018.

[13] Kartina, Linda dkk, "Efektivitas Sistem Temu Kembali Informasi *Online Public Access atalog (OPAC)* Dengan Tinjauan *Precision* Menggunakan Pendekatan Judul dan Subjek di Perpustakaan Universitas Muhammadiyah Bengkulu", Pustaloka : Jurnal Kajian Informasi dan Perpustakaan",Vol.11 No.2.2019.

[14] Nengsih, Warnia, "Analisa *Recall* dan *Precision* Menggunakan VSM pada *Text Mining*", InfoTekJar : Jurnal Nasional dan Teknologi Jaringan, Vol.5 No.1, 2020.