JIMP - Jurnal Informatika Merdeka Pasuruan
Vol 5 No 2 Agustus 2020
ISSN 2503-1945

Aplikasi Pencarian Bahan Pustaka Di Perpustakaan Menggunakan Metode Vector Space Model

Syaiful Bahri
Teknik Informatika, Universitas Widyagama, Malang
Email: syaifediting@gmail.com

Abstract

Library materials can be found by searching manually, but in addition to the possibility of missing relevant results and this method is also inefficient. Students also find it difficult to find library materials that suit their needs. So that the search engine is one of the solutions. This research uses the Vector Space Model (VSM) method with TF / IDF weighting on the top 10 documents as a way of ranking documents. There are 150 journals that are input using 4 selected queries by carrying out three stages of the process, namely the process documents, the Query process and finally the application of the Vector Space Model (VSM) method. In the next stage, this research calculates recall, precision and accuracy. The calculation result is on the precision of each query with a maximum value of 100% and the lowest value of 83%. Although there is a difference, it is not far. The recall value of all queries is 100%. Then on the results of the accuracy of all queries with a maximum value of 100% and a minimum of 96%. These results indicate that the IR system with the VSM method is effective and relevant to be used for searching library materials, also has a good and stable performance according to the trial database.

Keywords: Information Retrieval, Vector Space Model, TF / IDF, Precision Recall Accuracy

Abstrak

Bahan pustaka dapat ditemukan dengan melakukan pencarian secara manual, namun hal itu selain kemungkinan hasil yang relevan terlewatkan dan cara tersebut juga tidak efisien.Mahasiswa juga kesulitan untuk mencari bahan Pustaka yang sesuai kebutuhannya. Sehingga mesin pencarian merupakan salah satu solusinya Penelitian ini memanfaatkan metode Vektor Space Model (VSM) dengan pembobotan TF/IDF pada 10 dokumen teratas sebagai cara perangkingan dokumen.. Terdapat 150 jurnal yang di input dengan menggunakan 4 Query terpilih dengan melakukan tiga tahap proses yaitu proses dokumen, proses Query dan yang terahir adalah proses penerapan metode Vektor Space Model (VSM). Pada tahapan selanjutnya penelitian ini menghitung recall, Presisi dan Akurasi. Hasil penghitungan pada presisi dari masing-masing Query dengan nilai maksimal yaitu 100% dan nilai terendah 83%. Meskipun terdapat selisih namun tidak jauh. Nilai recall yang didapatkan dari semua Query adalah 100%. Kemudian pada hasil akurasi uji semua Query dengan nilai maksimal 100% dan 96% minimal. Hasil tersebut mengindikasikan bahwa system IR dengan metode VSM efektif dan relevan untuk digunakan untuk pencarian bahan Pustaka, juga memiliki performa yang baik dan stabil sesuai dengan database uji coba

Kata kunci: Information Retrieval, Vektor Space Model , TF/IDF, Presisi Recall Akurasi

I. PENDAHULUAN

Di era yang serba digital seperti sekarang ini muncul berbagai dampak baik yang positif maupun negative bagi kehidupan manusia, semua serba instan dan serba cepat, tak terkecuali dalam perkembangan ilmu pengetahuan dan perkembangan teknologi, semakin lama semuanya berkembang semakin cepat karena tahap hari ini akan dijadikan acuan dasar untuk selanjutnya esok hari.

Dinamika perubahan itupun datang dalam berbagai aspek, salah satunya yaitu dalam dunia Pendidikan khususnya dalam ilmu pengetahuan, teknologi dan informasi seperti yang sudah kita alami sendiri. Karena Pendidikan merupakan sesuatu yang berhubungan dengan budaya dan perdaban kehidupan manusia di perbagai penjuru dunia.

Perpustakaan merupakan jantung dari kampus dimana di dalamnya terdapat kumpulan koleksi, majalah, koran yang disusun berdasarkan sistem tertentu yang digunakan sebagai media dalam mencari ilmu dan wawasan bagi masyarakat.

Tujuan perpustakaan sendiri adalah sebagai pengarsipan yang ada diinstitusi adalah untuk

memelihara dokumen yang memiliki harga keilmuan yang nantinya dapat diakses dimasa depan sama mudahnya diakses pada saat ini. Dengan semakin kecilnya media data, penyimpanan menggunakan medai digital menjadi sangat menarik. Keuntungan penyimpanan menggunakan medai digital adalah pencarian dapat dilakukan tidak hanya di antara masukan katalog elektronik, tetapi dapat menampilkan sepanjang keseluruhan isi dokumen. Yang menjadi point penting dalam penerapan teknologi informasi dan komunikasi dalam pendidikan adalah penerapannya untuk menunjang peningakatan kwalitas pendidikan tersebut Pesatnya perkembangan teknologi informasi dan komunikasi mulai banyak dimanfaatkan dalam banyak bidang, salah satunya adalah bidang Pendidikan.(Safitri, 2018)

Tondeur et al (dalam Selwyn, 2011) menyatakan dalam penelitiannya bahwa teknologi digital kini sudah mulai digunakan di dalam lembaga pendidikan sebagai sarana untuk mendukung pembelajaran, baik sebagai alat informasi (yaitu sebagai sarana mengakses informasi) atau sebagai alat pembelajaran (yaitu sebagai sarana penunjang kegiatan belajar dan tugas). Teknologi merupakan hasil ciptaan manusia.

Universitas Widyagama merupakan salah satu perguruan tinggi swasta yang terletak ditengah Kota Malang, Jawa Timur, Indonesia. Universitas Widyagama didirikan oleh Yayasan Pembina Pendidikan Indonesia (YPPI) pada tahun 1971. Perpustakaan Universitas Widyagama berfungsi mendukung program akademik universitas yang tertuang dalam “Tridarma Perguruan Tinggi” yang mencakup pendidikan, penelitian dan pengabdian kepada masyarakat. Beberapa peran perpustakaan dalam hal mendukung proses pendidikan antara lain adalah memberikan informasi, mengkoordinasikan dan menggabungkan semua bentuk layanan untuk meningkatkan proses belajar mengajar, penelitian dan layanan umum. Pada akhirnya tujuannya adalah tercapainya proses peningkatan kualitas lulusan dalam hal pengembangan wawasan dan penguasaan keilmuannya. Perpustakaan UWG juga menyediakan Electronic Books (E-Books), Jurnal Internasional berbagai bidang pengetahuan, manual, dan lain sebagainya. Selain perpustakaan umum kampus juga terdapat perpustakaan di masing-masing fakultas, salah satunya di fakultas teknik informasi Pengelolaan data perpustakaan yang meliputi pengelolaan data buku, peminjaman dan pengembalian buku masih menggunakan cara konvensional, yaitu dengan menuliskannya ke dalam buku inventaris. Perpustakaan membutuhkan Sebuah system fungsi pencarian untuk menemukan Buku-Buku yang diinginkan. Proses pencarian dengan cara konvesional hanya menemukan Buku-Buku yang sesuai dengan kata kunci. Hal ini membuat proses pencarian menjadi kurang efektif, karena bisa saja pengguna tidak memasukkan kata kunci dengan tepat. Sementara Buku yang dicari tersedia dengan kata kunci berbeda namun masih dalam satu topik yang sama. Sistem temu kembali informasi merupakan bagian dari ilmu komputer yang berhubungan dengan pengambilan informasi dari dokumen-dokumen yang didasarkan pada isi dan konteks dari dokumen-dokumen itu sendiri. Proses dalam sistem temu kembali dapat digambarkan sebagai sebuah proses untuk mendapatkan dokumen yang relevan dari koleksi dokumen melalaui pencarian Query yang diimputkan user. Dengan berkembangnya teknologi dalam melakukan proses pencarian, maka untuk mendukung proses pencarian bahan pustaka pada aplikasi diimplementasikan dengan metode Vector Space Model.

Motode Vector Space Model adalah metode untuk melihat tingkat kedekatan atau kesamaan (smilarity) term dengan cara pembobotan term. Dokumen dipandang sebagi sebuah Vector yang memiliki magnitude (jarak) dan direction (arah). Pada Vector Space Model, sebuah istilah direpresentasikan dengan sebuah dimensi dari ruang Vector. Relevansi sebuah dokumen ke sebuah Query didasarkan pada similaritas diantara Vector dokumen dan Vector Query. (Ridwan & Sandi, 2019)

Penelitian terkait dengan metode Vector Space Model juga dilakukan antara lain Dr. Khalaf Khatatneh, M. Wedyan, DR. Mohamed Alham, DR Basem Alrifai (2005) dalam publikasinya yang berjudul ”Using new Data Structure to Implement Documents Vectors in Vector Space Model in Information Retrieval System” Bagaimana menggunakan tabel terstruktur dan Vector Space Model untuk lebih menghemat ruang untuk file dokumen yang sebelumnya memerlukan space besar dalam sistem temu kembali informasi. Aplikasinya dilakukan dengan menggunakan tabel dimana pada baris pertama digunakan kata kunci dan baris kedua digunakan pembobotan dari setiap kata kunci.

Solusi untuk mengatasi masalah di atas adalah dengan membuat sistem temu kembali informasi menggunakan metode Vector Space Model (VSM). Metode VSM dipilih karena cara kerja model ini efisien, mudah dalam representasi dan dapat diimplementasikan pada document-matching. (Ridwan & Sandi, 2019)

Algoritma Vector Space Model salah satu metode pencarian yang menghitung tingkat kemiripan antara kumpulan dokumen yang ada di basis data dengan dokumen yang dicari oleh pengguna. Oleh karena itu dengan menggunakan Algoritma Vector Space Model pada Pencarian Buku akan lebih efektif dibandingkan proses pencarian dengan cara konvensional akan lebih teliti karna pencarian dilakukan perkata dalam dokumen Perpustakaan

II. MASALAH

Dari latar belakang di atas maka dapat dikatakan bahwa rumusan masalah dalam penelitian ini adalah: bagaimana efekfivitas aplikasi Pencarian bahan pustaka di Perpustakaan fakultas Teknik Informatika Universitas Widyagama Malang dengan Menggunakan Metode Vector Space Model? Metode Vector Space Model digunakan hanya untuk proses pembobotan buku yang sesuai dengan kata kunci yang dicari. Untuk pembobotan Bahan Pustaka, Metode Vector Space Model membandingkan antara judul dokumen/jurnal dengan kata kunci/Query yang dicari pada sistem. Pengujian sistem akan dilakukan dengan 150 sampel judul jurnal dengan abstraknya, dengan uji hasil dengan presisi, recall dan akurasi.

III. LANDASAN TEORI

Information Retrival

Sistem Temu Kembali Informasi atau *InformationRetrieval* (IR) adalah menemukan materi (biasanya dokumen) dari sekumpulan data yang tidak terstruktur (biasanya teks) untuk memenuhi kebutuhan informasi dari koleksi yang besar (Manning et al., 2009).
Sistem temu kembali informasi merupakan suatu sistem yang menemukan (*retrieve)* informasi yang sesuai dengan kebutuhan *user* dari kumpulan informasi secara otomaris. Prinsip kerja sistem temu kembali informasi yaitu jika ada sebuah kumpulan dokumen dan seorang *user* yang memformulasikan sebuah pertanyaan (*request* atau *Query)* maka jawaban dari pertanyaan tersebut adalah sekumpulan dokumen yang relevan dan membuang dokumen yang tidak relevan (Gerald, 1988). Proses ini melibatkan berbagai tahapan dimulai dengan mewakili data dan diakhiri dengan mengembalikan informasi yang relevan kepada pengguna. Di antara tahap itu meliputi *filtering, searching, matching* dan perankingan. Tujuan utama dari sistem temu kembali informasi adalah untuk menemukan informasi yang relevan atau dokumen yang memenuhi kebutuhan informasi pengguna dari koleksi besar dokumen. Penggunaan *information retrieval* dapat diringkas sebagai berikut: memerlukan informasi dalam konteks aplikasi, pengguna memasukkan *Query* dengan harapan dapat mengambil satu sumber yang relevan. Dalam mencapai tujuan ini, *information retrieval* biasanya menerapkan 3 proses yaitu (Roshdi & Roohparvar, 2015):
1.Pada proses *indexing* dokumen direpresentasikan dalam bentuk konten yang Dirangkum
2. Pada proses *filtering* semua *stopwords* dan kata umum dihapuskan
3. *Searching* atau pencarian merupakan proses inti dari sistem temu kembali informasi banyak teknik untuk *retrieving* atau mengambil kembali dokumen yang sesuai dengan kebutuhan *user.*
Ada tiga dasar proses yang harus dalam sistem temu kembali informasi yaitu representasi dari konten dokumen, representasi kebutuhan informasi pengguna dan perbandingan dari dua representasi tersebut.
Representasi dokumen biasanya disebut indexing proses. Prosesnya berlangsung secara offline, yaitu pengguna akhir dari sistem pencarian informasi tidak terlibat secara langsung. Proses pengindeksan menghasilkan representasi dokumen. Proses representasi dari informasi yang dibutuhkan pengguna disebut sebagai proses perumusan *Query*. Representasi yang dihasilkan adalah *Query*. Membandingkan kedua representasi tersebut dikenal sebagai proses pencocokan. Proses ini akan menghasilkan daftar peringkat dokumen hasil pencarian. (Djoerd Hiemstra, 2009)

Komponen *Information Retrieval*

Information Retrieval dibuat dari sejumlah komponen perangkat lunak yang berkaitan dengan fungsi utama sistem, yaitu (Bates, 2012):
1. Menerima masukan dalam bentuk dokumen,mengekstrak informasi dari dokumen tersebut dan menyimpan informasi tersebut dalam bentuk yang dapat diakses dengan cepat agar sesuai dengan pencarian pengguna.
2. Menerima *Query* pengguna dan mengubah menjadi bentuk yang dapat dibandingkan dengan informasi tersimpan tentang dokumen.
Kedua proses ini sering disebut sebagai *indexing* dan *retrieval.* Bagian ini akan menjelaskan proses umum *indexing*, *retrieval* dan indeks untuk mencocokan *Query* dan dokumen.

Teks Processing

Proses indexing dalam sistem information retrieval berkaitan dengan pemberian representasi dokumen sesuai dengan peraturan dan proses yang ditetapkan untuk sistem tertentu. Salah satu bentuk yang paling sederhana adalah mengekstrak semua kata yang ada di setiap dokumen dan menyimpannya dalam indeks. Umumnya proses ini dikenal sebagai binary indexing yaitu sebuah kata tertentu yang terkait dengan dokumen. Normalisasi dokumen Proses indexing dalam sistem information retrieval berkaitan dengan pemberian representasi dokumen sesuai dengan peraturan dan proses yang ditetapkan untuk sistem tertentu. Salah satu bentuk yang paling sederhana adalah mengekstrak semua kata yang ada di setiap dokumen dan menyimpannya dalam indeks. Umumnya proses ini dikenal sebagai binary indexing yaitu sebuah kata tertentu yang terkait dengan dokumen. Normalisasi dokumen dan pembuatan index melibatkan text processing. Sistem dapat menggunakan berbagai proses ini (Bates, 2012). Ada beberapa tahapan umum yang digunakan pada proses text processing yaitu (Stefano Ceri, 2012):

a. *Document* Parsing.

Dokumen tersedia dalam berbagai Bahasa, rangkaian karakter dan format, seringkali dokumen yang sama berisi beberapa format dan Bahasa.

b. Lexical Analysis.

Lexical analysis yaitu proses tokenize sebuah dokumen menjadi kata. Menurut Croft et al. (2015) Tokenization merupakan proses pembentukan kata-kata dari urutan karakter dalam sebuah dokumen.. Semua huruf besar juga dikonversi menjadi huruf kecil.

c. *Stop-word Removal*

Stop-word Removal yaitu penghapusan kata yang berfrekuensi tinggi. Menurut Bates et al. (2012) Sistem *information retrieval* memiliki daftar kata-kata yang tidak berguna pada pencarian (contoh “an”, “the”, “a”). (Croft et al, 2015)

d. Filtering

Filtering adalah mengambil kata-kata penting dari hasil token. Bisa menggunakan algoritma *stoplist* (membuang kata yang kurang penting) atau *wordlist* (menyimpan kata penting).

e. *Filtering*

yaitu proses pembuangan *stopword* yang dimaksudkan untuk mengetahui suatu kata masuk kedalam *stopword* atau tidak. Pembuangan *stopword* adalah proses pembuangan *term* yang tidak memiliki arti atau tidak relevan

f. Stemming

Stemming yaitu proses yang bertujuan untuk menghilangkan akhiran kata untuk menormalkan kata. *Stemming* adalah proses untuk meminimalkan ketidakcocokan antara *Query* dan dokumen karena penggunaan bentuk kata yang berbeda (Peters, Braschler, & Clough, 2012). *Stemming* dapat mengurangi ukuran *indexing* sebanyak 40-50% (Vairaprakash Gurusamy, 2014)

g. Weighting

Kata-kata dalam sebuah teks memiliki kekuatn deskriptif yang berbeda.Oleh karena itu *term* indeks dapat diberi bobot yang berbeda untuk menghitung signifikannya dalam dokumen ataua koleksi dokumen.

h. Tokenisasi

Tokenisasi adalah pemotongan *string input* berdasarkan tiap kata yang menyusunnya. Pemecahan kalimat menjadi kata-kata tunggal dilakukan dengan menscan kalimat dengan pemisah (*delimiter*) *whitespace* (spasi, tab, dan *new line*). Secara garis besar *tokenisasi* adalah tahap memecah sekumpulan karakter dalam suatu teks kedalam satuan kata. Sekumpulan karakter tersebut dapat berupa karakter *whitespace*, seperti enter, tabulasi, spasi. (Susandi & Sholahudin, 2016)

Index dan *Query* Matching

Hasil *text processing* pada dokumen yang masuk disimpan dalam indeks sistem *information retrieval*. Indeks ini menyediakan mekanisme pencarian cepat untuk setiap token yang diekstraksi dengan pengolahan teks, bersama dengan informasi lainnya termasuk beberapa pengenal untuk dokumen yang berasal dari token, dan biasanya informasi lainnya. Struktur file spesifik yang digunakan untuk indeks bervariasi dari satu sistem ke sistem lainnya, namun yang paling umum adalah beberapa bentuk "*inverted file*". *Inverted file* menyediakan pemetaan antara *term* dan lokasi *occurrence*-nya pada koleksi teks. (Bates, 2012)

Porter Stemmer

Porter stemmer dikembangkan oleh Martin Porter pada tahun 1980 di University of Cambridge. *Porter stemmer adalah* algoritma akhiran penghapusan. Hal ini diketahui sangat sederhana dan ringkas. Algoritma ini didasarkan pada gagasan bahwa sebagian besar sufiks Bahasa Inggris terdiri dari yang sederhana lainnya. Dengan demikian, *stemmer* adalah algoritma linier yang menerapkan aturan morfologi secara berurutan sehingga

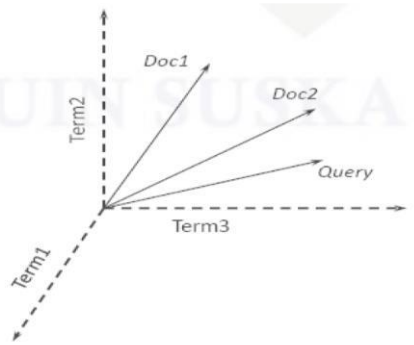
memungkinkan pelepasan sufiks secara bertahap. Secara khusus, algoritma ini memiliki lima langkah. Setiap langkah mendefinisikan seperangkat aturan. Untuk membendung sebuah kata, aturan itu diuji secara berurutan. Jika salah satu dari aturan ini sesuai dengan kata yang ada, maka syarat yang dilampirkan pada peraturan tersebut diuji. Begitu aturan diterima, akhiran akan dihapus dan langkah selanjutnya dijalankan. Jika aturan tersebut tidak diterima, maka peraturan berikutnya dalam langkah yang sama diuji, sampai salah satu peraturan dari langkah tersebut diterima atau tidak ada lagi peraturan dalam langkah tersebut dan oleh karena itu kontrol lolos ke langkah berikutnya. Proses ini berlanjut untuk semua enam langkah; Pada langkah terakhir *stem* yang dihasilkan dikembalikan (Nagamalai, 2009).

***Vector Space Model* (Model Ruang Vector)**

Vector Space Model (VSM) mempresentasikan setiap dokumen yang terdapat dalam database dan *Query* ke dalam Vector multidimensi. Dimensi dari Vector berkorespondensi dengan jumlah setiap term dalam database dan kumpulan term tersebut membentuk suatu ruang Vector. (Abdillah & Muktyas, 2018)

Salah satu model matematika yang digunakan pada sistem temu-kembali informasi untuk menentukan bahwa sebuah dokumen itu relevan terhadap sebuah informasi adalah *Vector Space Model* (VSM). Model ini akan menghitung derajat kesamaan antara setiap dokumen yang disimpan di dalam sistem dengan *Query* yang diberikan oleh pengguna. Model ini pertama kali diperkenalkan oleh Salton. *Vector Space Model* adalah suatu model yang digunakan untuk mengukur kemiripan antara suatu dokumen dengan suatu *Query*. (Susandi & Sholahudin, 2016)

Pada *Vector Space Model* dokumen dan *Query* diasumsikan sebagai *t* ruang dimensi Vector, dengan *t* sebagai indeks *term.* Sebuah dokumen *Di* direpresentasikan sebagai sebuah Vector dari indeks *term* (Croft, Metzler, & Strohman, 2015). Jika Vector dinormalisasi sehingga semua dokumen dan *Query* diwakili oleh Vector dengan panjang yang sama, kosinus sudut antara dua Vector identik adalah 1 (sudutnya nol), dan untuk dua Vector yang tidak memiliki nilai nol *term*, kosinus akan menjadi 0. Ukuran kosinus didefinisikan sebagai:



Gambar 1. Representasi Vector dokumen dan *Query* (Croft et al, 2015).

Perhitungan *Vector Space Model* menggunakan *cosine similarity* antara dokumen dan *Query* sebagai berikut

$$Cosine\ \mathrm{Di} = \frac{q \cdot d_j}{|q| * |d_j|}$$

Keterangan:
- $\boldsymbol{q}$: bobot *Query*
- $\boldsymbol{dj}$: bobot dokumen
- $|\boldsymbol{q}|$: akar jumlah kuadrat *q*
- $|\boldsymbol{dj}|$: akar jumlah kuadrat dokumen

Meskipun tidak ada definisi eksplisit relevansi pada *vector space model*, ada asumsi implisit bahwa relevansi terkait dengan kesamaan Vector *Query* dan dokumen. Dengan kata lain, dokumen "lebih dekat" dengan *Query* lebih mungkin relevan.

1. Pembobotan Istilah

Istilah di dalam suatu indeks harus bisa membedakan kepentingan dari sebuah dokumen pada sebuah informasi. Caranya yaitu dengan pemberian bobot kepada sebuah istilah terhadap suatu dokumen. Semakin tinggi bobot dari sebuah istilah maka semakin penting istilah tersebut dibandingkan dengan istilah lainnya di dalam sebuah dokumen. (Salton dan Buckley, 1987).

2. Pemfrofilan Text

Pemrofilan Teks (Kategorisasi teks atau klasifikasi teks) adalah sebuah pekerjaan dari bahasa dokumen yang alamiah untuk kategorisasi standar sesuai dengan kontennya [Sebastiani,2002]. Himpunan Kategori sering disebut “kosa kata terkontrol”. Pemrofilan teks atau kategorisasi teks adalah sebuah teknik yang lama dan bisa dikatakan sangat tradisional untuk pencarian informasi di perpustakaan

Metode TF/IDF

Faktor pembobotan untuk tiap kata dalam dokumen didefinisakan sebagai kombinasi term frequency dan invers document frequency. Untuk menghitung nilai bobot kata, digunakan rumus:

$$W_{in} = \frac{f_{in}}{log(k_n)}$$

Keterangan:

- Win : nilai bobot suatu term i terhadap sebuah dokumen
- fin : Nilai frekuensi term i didalam dokumen n
- Log(kn) : Jumlah term dalam dokumen n.

Indeks bobot *term* mencerminkan kepentingan relatif kata-kata dalam dokumen, dan digunakan dalam menghitung skor untuk perangkingan. Bentuk spesifik dari bobot ditentukan oleh model pengambilan. Komponen pembobotan menghitung bobot dengan menggunakan statistik dokumen dan menyimpannya dalam tabel pencarian. Bobot dapat dihitung sebagai bagian dari proses *Query*. Salah satu jenis yang paling umum digunakan pada model pencarian dikenal dengan pembobotan *TF-IDF (term frequency-inverse document frequency).* (W. Bruce Croft, Donald Metzler, 2015)

Term Frequency (TF) adalah metode pembobotan yang menentukan bobot dokumen berdasarkan pada kemunculan *term*. Lebih sering sebuah *term* muncul, bobot dokumen lebih tinggi untuk *term* tersebut. Hasil pembobotan ini selanjutnya akan digunakan dengan fungsi perbandingan untuk menentukan dokumen yang relevan. Inverse Document Frequency (IDF) meningkatkan nilai bobot dokumen berdasarkan rumus: "lebih banyak dokumen mengandung sebuah term, lebih kecil bobot term-nya, karena tidak dapat digunakan untuk membedakan relavan antara dokumen ". Rumus IDF adalah sebagai berikut (Cios, Pedrycz, Swiniarski, & Kurgan, 2007) :

$$\text{idf} = \log\frac{N}{dft} \dots\dots\dots\dots\dots\dots$$

Keterangan:

- N : total koleksi dokumen
- d : frekuensi dokumen

Penggabungkan definisi *term frequency* dan *inverse document frequency*, menghasilkan pembobotan untuk setiap *term* dalam setiap dokumen. Skema pembobotan *TF-IDF* yaitu (Cios et al., 2007)

Clustering

Clustering adalah model *data mining* yang diadopsi secara luas, sehingga data dibagi menjadi beberapa kelompok yang disebut *cluster* (Ilic, Rancic, & Spalevic, 2016). *Cluster* adalah daftar data yang memiliki karakteristik yang familier. Analisis *cluster* dapat dilakukan dengan mencari kesamaan antara data sesuai karakteristik yang terdapat pada data dan pengelompokan data yang sesuai kedalam *cluster* (Amandeep Kaur Mann, 2013).

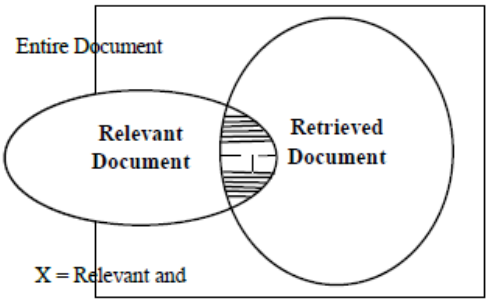
*Cluster*ing hasil pencarian telah dipelajari di bidang *Information Retrieval* Tujuan pengelompokan hasil pencarian adalah memberi pengguna gagasan tentang apa hasilnya. Ide ini berbentuk *cluster. Clustering* dalam konteks hasil pencarian berarti mengatur halaman hasil *Query* ke dalam kelompok berdasarkan kemiripannya satu sama lain. (Sadaf & Alam, 2012)

Suffix Tree

Suffix tree adalah struktur data yang berisi semua *suffix* dari *string* yang diberikan, sehingga bisa menjalankan banyak operasi *string* penting dengan lebih efisien. *String* bisa berupa *string* karakter atau *string* kata. *Suffix tree* untuk *string* S didefinisikan sebagai pohon sedemikian rupa sehingga: jalur dari *root* ke *leaves* memiliki hubungan satu lawan satu dengan akhiran S, semua tepi diberi label dengan *string* tidak kosong, semua *node internal* (kecuali *root)* memiliki setidaknya dua anak (Deng, 2012).

Akurasi *Information Retrieval*

Untuk mengetahui kualitas dari sistem temu kembali informasi dilakukan proses perhitungan akurasi. Sistem temu kembali informasi mengembalikan satu set dokumen sebagai jawaban atas *Query user.* Ada dua dokumen yang dapat ditemukan di *database* yaitu *relevant document* dan *retrieved document*. *Relevant document* adalah dokumen relevan dengan *Query user*. *Retrieved document* adalah dokumen yang diterima pengguna (Cios et al., 2007).



Gambar 2 Relasi antara Relevan dan Retrieved Document (Cios et al., 2007)

Ada dua ukuran dasar untuk menilai kualitas pengambilan teks yaitu (Gerald, 1988):

a. Precision

Precision adalah dokumen *retrieved* yang relevan dengan *Query*. Pengujian Ketepatan (Precision) ialah perbandingan jumlah dokumen relevan yang didapatkan sistem dengan jumlah seluruh dokumen yang terambil oleh sistem baik relevan maupun tidak relevan. Rumus *Precison* yaitu:

$$\text{Precision} = \frac{Number\ Of\ Relevant\ Items\ Retrieved}{Total\ Number\ Of\ Items\ Retrieved}$$

b. Recall

Recall adalah persentase dokumen yang relevan dengan *Query* pengguna yang terambil oleh sistem.

Recall adalah Pengujian Kelengkapan (Recall) ialah perbandingan jumlah dokumen relevan yang didapatkan sistem dengan jumlah seluruh dokumen relevan yang ada dalam koleksi dokumen (terambil ataupun tak terambil sistem).

$$\text{Recall} = \frac{Number\ Of\ Relevant\ Items\ Retrieved}{Total\ Number\ Of\ Relevant\ items\ in\ the\ collection}$$

Pengukuran presisi, recall dan akurasi perolehan informasi ini dilakukan agar bisa didapatkan kesimpulan dari pengujian yang dilakukan terhadap aplikasi dengan menggunakan metode yang ada.

IV. HASIL dan PEMBAHASAN

Rangking Hasil

Tahap ini user melakukan perangkingan dari data yang sudah didapatkan setelah melakukan klasterisasi dokumen dari masing-masing *Query* yang dicari. Ada 10 rangking teratas dari dokumen pada masing-masing *Query* yang dicari yang akan peneliti ambil pada penelitian ini.

Implementasi *Vector Space Model* dan Hasil

Penelitian ini peneliti menggunakan metode *Vector Space Model* dengan *tf/idf* dimana peneliti menggunakan 4 *Query* yang terdiri dari 1 kata (pendidikan) dan 2 kata (Mencari Nafkah) dan 3 kata (Jual Beli Online dan Sikap Percaya Diri) pada 150 dokumen jurnal yang diinput secara acak dan mengambil 10 dokumen teratas sesuai ranking dokumen dari masing-masing *Query*.

Tabel 1. Daftar Query

Token	*Query*
Q1	Pendidikan
Q2	Mencari Nafkah
Q3	Jual Beli Online
Q4	Sikap Percaya Diri

Hasil Evaluasi Pencarian

Terdapat beberapa dasar ukuran yang sering digunakan untuk mengetahui efektivitas sistem, untuk mengukur kemampuan hasil klasifikasi pada aplikasi VSM yang dibangun, sebuah model klasifikasi bisa dilihat dari nilai recall dan presisinya.

		Nilai sebenarnya	
		TRUE	FALSE
Nilai prediksi	TRUE	TP (True Positive) *Corect result*	FP (False Positive) *Unexpected result*
	FALSE	FN (False Negative) *Missing result*	TN (True Negative) *Corect absence of result*

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Rumus *Recall*, Pesisi da Akurasi

Presisi (P) adalah ukuran banyaknya dokumen yang ditemukan relevan, Precission menjawab pertanyaan “Berapa persen dokumen yang benar positif/sesuai dari keseluruhan dokumen yang diprediksi mucul?”dinyatakan dalam pecahan sebagai berikut;

$$presicion \quad : \frac{TP}{TP\ x\ FP} x100\%$$

sedangkan recall (R) adalah ukuran banyaknya dokumen yang relevan dapat ditemukan kembali, juga recall adalah rasio dari item yang relevan yang dipilih terhadap total jumlah item yang relevan. Nilai recall dapat diperoleh dari persamaan. Recall menjawab pertanyaan “Berapa persen dokumen yang diprediksi

muncul positif/true dibandingkan keseluruhan dokumen yang sebenarnya mucul/true"dinyatakan dalam pecahan sebagai berikut;

$$Recall \quad : \frac{TP}{TP\ x\ FN} x100\%$$

Accuracy

Merupakan rasio prediksi Benar (positif dan negatif) dengan keseluruhan data. Akurasi menjawab pertanyaan "Berapa persen dokumen yang benar diprediksi sesuai dan Tidak sesuai dari kesuluruhan dokumen yang di input oleh user"

$$Accuracy \quad : \frac{TP+TN}{TP+TN+FP+FN} x100\%$$

Tabel 2. Hasil Presisi, *Recall dan* Akurasi pada *Query* Pendidikan

Pendidikan		Nilai sebenarnya	
		True	False
Nilai Prediksi	True	38	0
	False	0	112

$$Precission \quad : \frac{38}{38\ x\ 0} x100\%$$

$$= 1\ x100 = 100\%$$

$$Recall \quad : \frac{38}{38\ x\ 0} x100\%$$

$$= 1\ x100 = 100\%$$

$$Accuracy \quad : \frac{38+112}{38+112+0+0} x100\%$$

$$\frac{150}{150} x100\%$$

$$= 1\ x\ 100 = 100\%$$

$$= 100\%$$

Tabel 3. Hasil Presisi, *Recall dan* Akurasi pada *Query* Mencari Nafkah

Mencari Nafkah		Nilai sebenarnya	
		True	False
Nilai Prediksi	True	25	5
	False	0	120

$$Precission \quad : \frac{25}{25\ x\ 5} x100\%$$

$$= 0{,}83\ x100\% = 83\%$$

$$Recall \quad : \frac{25}{25\ x\ 0} x100\%$$

$$= 1\ x\ 100\% = 100\%$$

$$Accuracy \quad : \frac{25+120}{25+120+5+0} x100\%$$

$$\frac{145}{150} x100\%$$

$$= 0{,}96\ x\ 100 = 96\%$$

$$= 96\%$$

Tabel 4. Hasil Presisi, *Recall dan* Akurasi pada *Query* Jual Beli *Online*

Jual Beli Online		Nilai sebenarnya	
		True	False
Nilai Prediksi	True	25	3
	False	0	112

$$Precission \quad : \frac{25}{25\ x\ 3} x100\%$$

$$= 0{,}89\ x100\% = 89\%$$

$$Recall \quad : \frac{25}{25\ x\ 0} x100\%$$

$$= 1\ x\ 100\% = 100\%$$

$$Accuracy \quad : \frac{25+122}{25+122+3+0} x100\%$$

$$\frac{147}{150} x100\%$$

$$= 0{,}98\ x\ 100 = 98\%$$

$$= 98\%$$

Tabel 5. Hasil Presisi, *Recall dan* Akurasi pada *Query* Sikap Percaya diri

Sikap Percaya Diri		Nilai sebenarnya	
		True	False
Nilai Prediksi	True	25	4
	False	0	121

$$Precission : \frac{25}{25 x 4} x100\%$$

$$= 0{,}86 \text{ x}100\% = 86\%$$

$$Recall : \frac{25}{25 x 0} x100\%$$

$$= 1 \text{ x } 100\% = 100\%$$

$$Accuracy : \frac{25+121}{25+120+5+0} x100\%$$

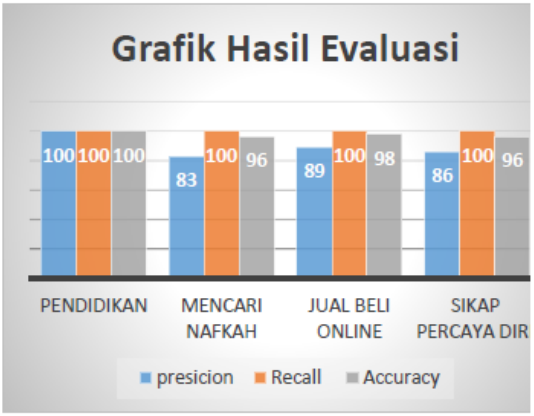
$$\frac{146}{150} x100\%$$

$$= 0{,}97 \text{ x } 100 = 97\%$$

$$= 96\%$$

Tabel 6. Hasil Presisi, *Recall dan* Akurasi pada keseluruhan *Query*

No	*Query*	Presisi	Recall	Akurasi
1.	Pendidikan	100%	100%	100%
2.	Mencari Nafkah	83%	100%	96%
3.	Jual Beli Online	89%	100%	98%
4.	Sikap Percaya Diri	86%	100%	97%



Hasil dari uji coba penelitian ini dapat dilihat dalam tabel bahwa secara umum aplikasi yang dibangun dapat membedakan dengan baik antara 4 *Query* yang di uji cobakan, yaitu Pendidikan, jual beli online, mencari nafkah, dan sikap percaya diri, dibuktikan dengan rata-rata tingkat presisi, *recall* dan akurasi yang baik.

Hasil precission diperoleh dari rumus

$$\frac{TP}{TP \; x \; FP} x100\%$$

Suatu sistem temu kembali dinyatakan efektif apabila hasil penelusuran mampu menunjukkan ketepatan (precision) yang tinggi sekalipun perolehannya rendah (Rowley, 1992).

Pengukuran efektivitas suatu sistem temu kembali dapat dilakukan dengan perhitungan terhadap nilai perolehan (recall), nilai ketepatan (precision), dan jatuhan semu (fallout) (Tague-Sutcliffe, 1992; Conlon dan Conlon, 1996). Namun diantara metode tersebut, perhitungan ketepatan merupakan cara yang paling umum digunakan (Su, 1992; Tague-Sutcliffe, 1992) Menurut Rowley dalam Hasugian (2003: 5), suatu sistem temu kembali informasi dinyatakan efektif apabila hasil penelusuran mampu menunjukkan ketepatan (precision) yang tinggi sekalipun perolehannya rendah. Dari hasil keseluruhan hal ini berarti bahwa model pencarian yang dibangun dengan metode SVM ini memiliki performa yang baik dan stabil untuk semua database coba. Relevansi system pencarian ini dapat dilihat dari hasil presisi (ketetapan) dengan nilai minimum 83% hingga nilai maksimum 100%. Nilai dari masing-masing *Query* dapat dilihat pada tabel di atas, dimana semua recall dari 4 *Query* adalah 100%, Sedangkan hasil dari uji akurasi dari 4 *Query* dengan nilai minimum 96% dan nilai maksimum 100%. Dapat dinyatakan bahwa system pencarian dengan metode VSM tersebut efektif digunkaan sebagai motode pencarian bahan oustaka diperpustakaan jurusan IT di Universitas Widyagama Malang sehingga diharapkan motode ini dapat membantu mahasiswa yang lain dalam mencari dokumen atau bahan Pustaka yang diinginkan.

V. KESIMPULAN

Pada uji coba ini menggunakan 4 kata kunci (*Query*) dengan 150 dokumen yang dinput dan disesuaikan dengan *Query* pengguna.

Suatu system temu Kembali informasi dikatakan efektif apabila hasil penelusuran mampu menunjukkan ketepatan (presisi) yang tinggi sekalipun perolehannya rendah. Dalam penelitian ini hasil presisi yang diperoleh adalah 100% - 83% dan hasil ini termasuk kategori tinggi. Adapun nilai recall dari semua *Query* adalah 100% dengan hasil akurasi 100% - 96%. Dari data tersebut dapat dinayatakan bahwa dokumen yang dicari muncul dengan tingkat presisi, *recall* dan akurasi yang tinggi artinya dokumen yang diharapkan muncul sesuai yang di harapkan. Maka system temu Kembali informasi dengan metode *Vector Space Model* efektif untuk digunakan sebagai system pencarian bahan Pustaka diperpustakaan. Secara keseluruhan hasil uji coba semua *Query* yang digunakan dapat dinyatakan bahwa metode VSM ini memiliki performa yang baik dan stabil sesuai dengan database uji coba.

VI. SARAN

Adapun saran untuk penelitian berikutnya adalah diharapkan agar bisa mengoptimalisasikan pada tahap pemilihan

Query, juga bisa memilih kata kunci yang lebih spesifik sehingga relevansi hasil *Retrieve* juga semakin tinggi, sehingga

DAFTAR PUSTAKA

- Safitri, Cici Erza. (2018). ("*Implementasi Metode Vector Space Model Dan Suffix Tree Clustering Pada Sistem Temu Kembali Informasi Dan Clustering E-Jurnal"),* Skripsi, Fakultas Sains Dan Teknologi, Jurusan Informatika, Universitas Islam Negeri Sultan Syarif Kasim, Riau.
- putubuku, "Recall & Precision," Ilmu Perpustakaan & Informasi – diskusi dan ulasan ringkas, 27-Mar-2008. [Online]. Available: http://iperpin.wordpress.com/2008/03/27/recall-precision/. [Accessed: 16-Jun-2013].
- "Precision and recall," Wikipedia – The Free Encyclopedia. [Online]. Available: http://en.wikipedia.org/wiki/Precision_and_recall. [Accessed: 16-Jun-2013].
- "Accuracy and precision," Wikipedia – The Free Encyclopedia. [Online]. Available: https://en.wikipedia.org/wiki/Accuracy_and_precision. [Accessed: 16-Jun-2013].
- B. Raharjo, "Presisi Dan Akurasi," Beni Raharjo – Nature, Environment, Remote Sensing, GIS, IT and Myself, 17-Mar-2011. [Online]. Available: http://www.raharjo.org/math/presisi-dan-akurasi.html. [Accessed: 16-Jun-2013].\
- DATAQ, "Perbedaan: precision, recall & accuracy", https://dataq.wordpress.com/2013/06/16/perbedaan-precision-recall-accuracy/. 16 juni 2013.
- Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval*. *Cambridge University Press*. https://doi.org/`Gerald, S. (1988). *Automatic text processing*. Addison Wesley Publishing Company.
- Roshdi, A., & Roohparvar, A. (2015). Review : Information Retrieval Techniques and Applications. *International Journal Of Computer Networks and Communications Security*, *3*(9), 373–377.
- Djoerd Hiemstra. (2009). Information Retrieval Models. *Information Retrieval: Searching in the 21st Century.*
- Saini, C. V. A. (2016). Information Retrieval in Web Crawling : A Survey. *International Conference On Advances in Computing, Communications Adn Informatics (ICACCI)*.
- Bates, M. J. (2012). *Understanding Information Retrieval System*. CRC Press.
- Stefano Ceri, A. B. (2012). Web information retrieval (pp. 83–110). Springer. https://doi.org/10.1201/b12118-8
- W. Bruce Croft, Donald Metzler, T. S. (2015). *Information Retrieval in Practice*. Pearson Education, Inc
- Vairaprakash Gurusamy, S. K. (2014). Preprocessing Techniques for Text Mining.
- Porter, M. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.
- Croft, W. B., Metzler, D., & Strohman, T. (2015). *Information Retrieval in Practice*. *Pearson Education, Inc.*
- W. Bruce Croft, Donald Metzler, T. S. (2015). *Information Retrieval in Practice*. Pearson Education, Inc.
- Cios, K. J., Pedrycz, W., Swiniarski, R. W., & Kurgan, L. A. (2007). *Data Mining A Knowledge Dizcovery Approach*.
- Amandeep Kaur Mann, N. K. (2013). Review paper on clustering techniques. *Global Journal of Computer Science and Technology Software & Data Engineering*, *23*(5).
- Ilic, M., Rancic, D., & Spalevic, P. (2016). Comparison of Data Mining Algorithms, Inverted Index Search and Suffix Tree Clustering Search. *Facta Universitatis, Series: Automatic Control and Robotics*, *15*(3), 171.
- Sadaf, K., & Alam, M. (2012). WEB SEARCH RESULT C LUSTERING – A REVIEW. *International Journal of Computer Science & Engineering Survey (IJCSES)*, *3*(4), 85–92.
- Deng, F. (2012). *Web Service Matching based on Semantic Classification Title: Web Service Matching based on Semantic Classification*. *School of Health and Society Department Design and Computer Science Kristianstad University*.
- Wang, D., Liu, L., Dong, J., & Zheng, J. (2015). Search results clustering algorithm based on the suffix tree. *Proceedings - 2015 2nd International Conference on Information Science and Control Engineering, ICISCE 2015*, 456–460.
- Gerald, S. (1988). *Automatic text processing*. Addison Wesley Publishing Company.
- Susandi, Sholahudin. (2016). ("Pemanfaatan Vector Space Model pada Penerapan Algoritma Nazief Adriani, KNN dan Fungsi *Similarity Cosine* untuk Pembobotan IDF dan WIDF pada *Prototipe*

Sistem Klasifikasi Teks Bahasa Indonesia"), Jurnal ProTekInfo, Vol. 3,

- Herliani, ("Aplikasi Pencarian Buku Dengan Menggunakan Metode Tf/Idf Dan *Vector Space Model* Berbasis Web Pada Perpustakaan Sekolah Menengah Atas Negeri 2 Pangkalpinang").
- Mutia Ayudita1, Pandu Adikara (2018). Sistem Pencarian Jurnal Ilmiah Cross Language dengan Metode Vector Space Model (VSM). (2). 6839-6840
- Ridwan**,** Armawan Sandi2, (2019). Penerapan Mesin Pencari Informasi Dengan Menggunakan Metode *Vector Space Model.* (3).