SISTEM TEMU KEMBALI INDEX BERITA MENGGUNAKAN VECTOR SPACE MODEL

Afri Yudha

afri.yudha@ibm.ac.id

Program Studi Teknik Informatika Institut Bisnis Muhammadiyah Bekasi

ABSTRACT

Searching with its high correctness becomes very crucial because documents maintained are in big data. Vector space model is the model to measure the similarities between documents and queries. This journal contains searching system news index for example with sports and politics which exist in portal web. By using vector space model, the searching results precisely 67% recall 100% and its accuracy 67% in K.

***Key Words**: News Index, Vector Space Model, Consine Similarities, tf-idf, precision-recall.*

1. PENDAHULUAN

Index berita merupakan suatu hal yang penting dalam sebuah media informasi saat ini, di mana hal ini sangat membantu dalam kemudahan pencarian berita yang diinginkan. Pencarian *index* berita saat ini dilakukan dengan menggunakan mesin pencari atau *system*.

Sistem Temu Kembali Informasi (STKI), *user* menuliskan *query* dan mesin pencari akan menampilkan hasil pencarian. Mesin pencari yang sudah ada dan banyak digunakan saat ini memberikan hasil perolehan pencarian yang banyak (banyak dokumen yang terambil), sehingga diperlukan waktu untuk menentukan hasil pencarian yang relevan. Menentukan hasil yang relevan sesuai dengan keinginan *user* dengan jumlah hasil pencarian yang banyak akan menyulitkan *user*. Hal ini terjadi karena dokumen yang terambil oleh sistem jumlahnya banyak, maka sistem berkemungkinan menampilkan hasil pencarian yang tidak relevan. Banyaknya dokumen hasil pencarian ini membuat waktu yang dibutuhkan dalam pencarian menjadi lebih banyak dari yang diharapkan.

Menurut Putu Laxman Pendit. [et al.] (2007), sistem temu kembali informasi atau *information retrieval* (IR) merujuk pada keseluruhan kegiatan yang meliputi pembuatan wakil informasi (*representation*), penyimpanan (*storage*), pengaturan (*organization*) sampai ke pengambilan (*access*).

Lee Pao (1989) mengatakan, dalam sistem temu kembali informasi memiliki prinsip ketepatan dalam menemukan informasi yang diperlukan yaitu *recall* dan *precision*. Menurut Lee Pao, ada dua hal penting yang biasanya digunakan dalam mengukur kemampuan suatu sistem temu kembali informasi yaitu rasio atau perbandingan dari perolehan (*recall*), dan ketepatan (*precision*).

Seiring dengan perkembangan teknologi seperti sekarang, membuat alat penelusuran informasi menjadi semakin modern dan canggih ditambah lagi dengan

sistem informasi yang memudahkan pengguna di dalam temu kembali informasi menggunakan OPAC.

Menurut Supriyanto dalam Siti Febrianti (2016) *Online Public Acces Catalogue* (OPAC) adalah sebuah fitur yang digunakan untuk memfasilitasi pengunjung *web* untuk mencari katalog koleksi perpustakaan yang dapat diakses oleh umum yang fungsinya sama dengan kartu katalog yang tersedia di perpustakaan pada umumnya. Katalog *online* ini dapat dimanfaatkan dengan sangat mudah sebagai bibliografi atau bahkan indeks pun terdapat pada katalog *online*. Berdasarkan penjelasan di atas bahwa OPAC merupakan kumpulan katalog *online* yang dapat memudahkan pengguna mengakses informasi secara mudah.

Perkembangan penelusuran informasi saat ini menghasilkan *recall* yang tinggi dan *precision* yang rendah. *Recall* yang tinggi diartikan bahwa dokumen yang dihasilkan dalam penelusuran dokumen adalah banyak, sedangkan *precision* rendah dapat diartikan bahwa dokumen yang diharapkan dapat ditemukan sedikit.

2. METODE

Dalam penulisan karya ilmiah ini, penulis menggunakan metode penelitian *deskriptif kualitatif* dengan menggunakan studi kasus. Menurut Arikunto (2006) mengatakan bahwa penelitian deskriptif kualitatif adalah prosedur atau cara memecahan masalah penelitian dengan memaparkan keadaan objek yang diselidiki seperti (seseorang, lembaga, pabrik, dan lain-lain) sebagaimana adanya berdasarkan fakta- fakta yang aktual pada saat sekarang atau pada saat penelitian dilakukan. Arikunto (2006) juga berpendapat bahwa dalam penelitian deskriptif pada umumnya tidak perlu merumuskan hipotesis.

Dalam penelitian ini, penulis melakukan pencarian data dari *web* portal *detik.com* rentang waktu 30 September 2017 di mana sampel yang digunakan sebanyak 7 data judul dari berita politik dan olahraga.

A. Sistem Temu Kembali Informasi

Sistem temu kembali informasi merupakan suatu sistem yang menemukan *(retrieve)* informasi yang sesuai dengan kebutuhan *user* dari kumpulan informasi secara otomatis. Prinsip kerja sistem temu kembali informasi, jika ada sebuah kumpulan dokumen dari seorang *user* yang memformulasikan sebuah pertanyaan (*request* atau *query*). Jawaban dari pertanyaan tersebut adalah sekumpulan dokumen yang relevan dan membuang dokumen yang tidak relevan (Salton, 1989).

Sistem temu kembali informasi akan mengambil salah satu dari kemungkinan tersebut. Sistem temu kembali informasi dibagi dalam dua komponen utama yaitu sistem pengindeksan *(indexing)* menghasilkan basis data sistim dan temu kembali merupakan gabungan dari *user interface* dan *look-up-table*. Sistem temu kembali informasi didesain untuk menemukan dokumen atau informasi yang diperlukan oleh *user*. Fungsi utama sistem temu kembali informasi (Salton, 1989) adalah:

a) Mengidentifikasi sumber informasi yang relevan dengan minat masyarakat pengguna yang ditargetkan.
b) Menganalisis isi sumber informasi (dokumen).
c) Merepresentasikan isi sumber informasi dengan cara tertentu yang memungkinkan untuk dipertemukan dengan pertanyaan pengguna.
d) Merepresentasikan pertanyaan (*query*) *user* dengan cara tertentu yang memungkinkan untuk dipertemukan sumber informasi yang terdapat dalam basis data.
e) Mempertemukan pernyataan pencarian dengan data yang

tersimpan dalam basis data.

f) Menemu-kembalikan informasi yang relevan.

g) Menyempurnakan unjuk kerja sistem berdasarkan umpan balik yang diberikan oleh *user*.

B. Metode TF/IDF

Metode TF-IDF (*Term Frequncy Inverse Document Frequency*) merupakan suatu cara untuk memberikan bobot hubungan suatu kata (*term*) terhadap dokumen (Robertson, 2005). Metode ini menggabungkan dua konsep untuk perhitungan bobot yaitu, frekuensi kemunculan sebuah kata di dalam sebuah dokumen tertentu dan inverse *frekuensi* dokumen yang mengandung kata tersebut. Terdapat beberapa cara atau metode dalam melakukan pembobotan kata pada metode TF-IDF, yaitu melalui skema pembobotan *query* dan dokumen. Skema pembobotan *query* dan dokumen merupakan salah satu jenis pembobotan kata pada *System for the Mechanical Analysis and Retrieval of Text (*SMART) atau sering disebut sebagai SMART *notation* (notasi SMART). Pada notasi SMART, merepresentasikan pembobotan ke dalam bentuk *ddd.qqq* (Manning *et al,* 2009).

C. Tokenisasi

Tokenisasi merupakan proses pemisahan suatu rangkaian karakter berdasarkan karakter spasi, dan mungkin pada waktu yang bersamaan dilakukan juga proses penghapusan karakter tertentu, seperti tanda baca. Sebagai contoh, kata-kata "*computer*", "*computing*", dan "*compute*". Semua berasal dari *term* yang sama yaitu "*comput*", tanpa pengetahuan sebelumnya dari morfologi bahasa Inggris. Token seringkali disebut sebagai istilah (*term*) atau kata, sebagai contoh sebuah token merupakan suatu urutan karakter dari dokumen tertentu yang dikelompokkan sebagai unit semantik yang berguna untuk diproses (Salton, 1989).

D. Vector Space Model

Vector Space Model (VSM) adalah metode untuk melihat tingkat kedekatan atau kesamaan *(similarity) term* dengan cara pembobotan *term*. Dokumen dipandang sebagai sebuah vektor yang memiliki *magnitude* (jarak) dan *direction* (arah). Pada *Vector Space Model,* sebuah istilah direpresentasikan dengan sebuah dimensi dari ruang vektor. Relevansi sebuah dokumen ke sebuah *query* didasarkan pada similaritas di antara vektor dokumen dan vektor *query* (Baeza, 1999).

VSM memberikan sebuah kerangka pencocokan parsial adalah mungkin. Hal ini dicapai dengan menetapkan bobot non-biner untuk istilah indeks dalam *query* dan dokumen. Bobot istilah yang akhirnya digunakan untuk menghitung tingkat kesamaan antara setiap dokumen yang tersimpan dalam sistem dan permintaan *user*.

Dokumen yang terambil disortir dalam urutan yang memiliki kemiripan, model vektor memperhitungkan pertimbangan dokumen yang relevan dengan permintaan *user*. Hasilnya adalah himpunan dokumen yang terambil jauh lebih akurat (dalam arti sesuai dengan informasi yang dibutuhkan oleh *user*).

Dalam VSM koleksi dokumen direpresentasikan sebagai sebuah matrik *term document* (atau matrik *term frequency*). Setiap sel dalam matrik bersesuaian dengan bobot yang diberikan dari suatu *term* dalam dokumen yang ditentukan. Nilai nol berarti bahwa *term* tersebut tidak ada dalam dokumen.

E. Precision Dan Recall

Perhitungan Precision dan *Recall* merupakan perhitungan yang umum digunakan untuk mengevaluasi hasil pencarian. *Recall* adalah perbandingan antara hasil pencarian yang relevan dengan seluruh data relevan yang ada pada koleksi *database*. Sedangkan *precision* adalah perbandingan antara hasil pencarian yang

relevan terhadap semua pencarian yang berhasil di retrieve. Recall dan precision digunakan dalam mengukur tingkatkeberhasilan pencarian. Semakin tinggi ukuran *precision* dan *recall*-nya maka semakin bagus strategi pencariannya. Selain itu, suatu sistem dinyatakan efektif apabila hasil penelusuran mampu menunjukkan ketepatan (*precision*) yang tinggi.

$$\text{Precision} = \frac{tp}{tp + fp} \qquad \text{Recall} = \frac{tp}{tp + fn}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \qquad \text{True Negative Rate} = \frac{tn}{tn + fp}$$

Gambar 1. Rumus dan Precision dan Recall

F. Consine Similarity

Secara umum, fungsi *similarity* adalah fungsi yang menerima dua buah objek dan mengembalikan nilai kemiripan (*similarity*) antara kedua objek tersebut berupa bilangan riil. Umumnya, nilai yang dihasilkan oleh fungsi *similarity* berkisar pada interval [0...1]. Namun ada juga beberapa fungsi *similarity* yang menghasilkan nilai yang berada di luar interval tersebut. Untuk memetakan hasil fungsi tersebut pada interval [0...1] dapat dilakukan normalisasi.

Cosine similarity adalah perhitungan kesamaan antara dua vektor *n* dimensi dengan mencari kosinus dari sudut di antara keduanya dan sering digunakan untuk membandingkan dokumen dalam *text mining*.

3. HASIL DAN PEMBAHASAN

Dalam melakukan implementasi vector space model terhadap metode pembobotan TF-IDF pada suatu sistem temu kembali, penulis melakukan random pada *indexing* berita yang terdapat di *website* www.detik.com pada tanggal 30 September 2017 dengan menggunakan 7 sampel judul berita.

ID	JUDUL	KATEGORI
1	Novanto Menang di Praperadilan, Keluarga Bersyukur	Politik
2	Empat Balapan Tersisa yang Masih Misteri untuk Marquez	Olahraga
3	Pemenang di Marina Bay Biasanya Akan Jadi Juara Dunia	Olahraga
4	Risma Sepakat dengan Megawati, Pilkada Bukan Kalah Menang	Politik
5	Lelang Mobil KPK, Pria Ini Menangi Alphard Fuad Amin Rp 301 Juta	Politik
6	Novanto Menang Praperadilan, Status Tersangka Dinyatakan Tak Sah	Politik
7	Balapan di Sepang, Hamilton Yakin Menang?	Olahraga

Tabel 1. Sampel Berita Sumber Detik.Com

Sampel yang telah didapatkan, akan dibuat *indexer* dengan *term-term* yang berasal dari *sample* atau data training yang ada.

JUDUL	ID						
	1	2	3	4	5	6	7
adil	1					1	
alphard					1		
balap		1					1
biasa			1				
bukan				1			
dunia			1				
empat		1					
fuad amin					1		
hamilton							1
jadi			1				
juara			1				
juta					1		
kalah				1			
keluarga	1						
kpk					1		
lelang					1		
marina bay			1				
marquez		1					
masih		1					
megawati				1			
menang	1		1	1	1	1	1
misteri		1					
mobil					1		
novanto	1					1	
nyata						1	
pilkada				1			
pria					1		
risma				1			
rupiah					1		
sah						1	
sangka						1	
sepakat				1			
sepang							1
sisa		1					
status						1	
syukur	1						
yakin							1

Gambar 2. Hasil Indexer

Lalu dilakukan perhitungan TF dan IDF, maka akan didapatkan hasil seperti gambar di bawah ini:

.

JUDUL	TF	DF/TF	IDF
adil	2	3,5	0,54406804
alphard	1	7	0,84509804
balap	2	3,5	0,54406804
biasa	1	7	0,84509804
bukan	1	7	0,84509804
dunia	1	7	0,84509804
empat	1	7	0,84509804
fuad amin	1	7	0,84509804
hamilton	1	7	0,84509804
jadi	1	7	0,84509804
juara	1	7	0,84509804
juta	1	7	0,84509804
kalah	1	7	0,84509804
keluarga	1	7	0,84509804
kpk	1	7	0,84509804
lelang	1	7	0,84509804
marina bay	1	7	0,84509804
marquez	1	7	0,84509804
masih	1	7	0,84509804
megawati	1	7	0,84509804
menang	6	1,1667	0,06694679
misteri	1	7	0,84509804
mobil	1	7	0,84509804

novanto	2	3,5	0,54406804
nyata	1	7	0,84509804
pilkada	1	7	0,84509804
pria	1	7	0,84509804
risma	1	7	0,84509804
rupiah	1	7	0,84509804
sah	1	7	0,84509804
sangka	1	7	0,84509804
sepakat	1	7	0,84509804
sepang	1	7	0,84509804
sisa	1	7	0,84509804
status	1	7	0,84509804
syukur	1	7	0,84509804
yakin	1	7	0,84509804

Gambar 3. Hasil Perhitungan TF-IDF

Setelah penulis mendapatkan hasil perhitingan TF-IDF maka penulis melakukan perhitungan dengan *consine similarity* dan didapatkan hasil seperti gambar di bawah ini:

Consine	Jarak	Katego
Consine Similarity (Q,D5)	0,298362	Politik
Consine Similarity (Q,D7)	0,190854	Olahraga
Consine Similarity (Q,D2)	0,105784	Olahraga
Consine Similarity (Q,D1)	0,003127	Politik
Consine Similarity (Q,D6)	0,002394	Politik
Consine Similarity (Q,D3)	0,002353	Olahraga
Consine Similarity (Q,D4)	0,002148	Politik

Gambar 4 Hasil Consine Similarity

Setelah mendapati perhitungan *consine similarity* maka penulis mencari nilai perhitungan *precission* dan *recall*, di mana hasilnya dapat dilihat seperti gambar dibawah ini.

Precision	TP/TP+FP
	=2/2+1
	66,67%
Recall	TP/TP+FN
	=2/2+0
	100%

Gambar 5 hasil perhitungan recall dan precission

Dengan diperolehnya nilai *Consine Similarity* maka didapatkan nilai tertinggi dengan *query* "menang balapan mobil" menghasilkan kategori 2 olahraga dan 1 politik, dengan tingkat *Precision* 66.7% dan *Recall* 100%.

4. KESIMPULAN

Berdasarkan hasil perhitungan sistem temu kembali *Index* berita dengan menggunakan *vector space model* didaptakan kesimpulan:

1. Presisi dan *recall* yang dihasilkan dengan rumus *consine similarity* menghasilkan nilai sebesar 67% untuk presisi dan 100% untuk *recall*.
2. Nilai akurasi tergantung pada banyaknya data training yang ada, semakin banyak data training yang ada, semakin tinggi nilai akurasinya.

Selain itu penulis dapat menggunakan hasil penelitian ini untuk diimplementasikan dalam sebuah aplikasi dengan penambahan data training yang lebih banyak lagi, mengingat banyaknya *Index* berita yang terdapat di setiap media informasi.

DAFTAR PUSTAKA

Arikunto, S. (2006). *Prosedur penelitian: Suatu pendektan praktek*. Jakarta: Rineka Cipta.

Baeza R.Y, Neto R. (1999). *Modern Information Retrieval, Addison' Manning, Christopher D, Prabhakar Raghavan dan Hinrich Schutze. 2009. An Introduction to Information Retreival, England*: Cambridge University Press.

Pendit, P. L. (2008). *Perpustakaan Digital Dari A Sampai Z*. Jakarta: Cita karyakarsa.

Pendit, P.L. [et al.]. (2007). *Perpustakaan Digital; Perpustakaan Digital: Perpekstif Perpustakaan Perguruan Tinggi Indonesia*, Jakarta: Sagung Seto.