| *Journal Computer Science and Information Technology(JCoInT)*
Program Studi Teknologi Informasi, Fakultas Sains & Teknologi, Universitas Labuhanbatu
Volume: 3 Issue 2, 05,2022, Page: 73-79
eISSN: 2774-7751
Journal Homepage : http://jurnal.ulb.ac.id/index.php/JCoInT/index

Information Text Retrieval Untuk Pencarian Data Penilaian Mengacu Pada Saran Dari Pengunjung Menggunakan *Vector Space Modelimplementasi*

Khoirunsyah Dalimunthe[1], B.Herawan Hayadi[2]

Email: choyrun@gmail.com[1], b.herawan.hayadi@gmail.com[2]

[1,2] Ilmu Komputer, Fakultas Teknik dan Informatika, Universitas Potensi Utama
Correspondence: E-mail: aliakbarritonga@gmail.com

A B S T R A C T S

Dengan perkembangan zaman yang begitu pesat dan informasi yang terus berkembang maka dibutuhkan suatu sistem otomatis untuk membantu kita dalam mempermudah menemukan informasi yang ingin kita ketahui secara cepat. Salah satu metode dalam mencari informasi retrieval adalah dengan menggunakan metode vector space model. Metode pengolahan data yang akan dilakukan pada penelitian ini terbagi menjadi dua, yaitu pengolahan data offline dan real time. implementasi information retrieval saat proses preprocessing dokumen dan pembobotan tf-idf adalah me-retrieve dokumen atau menemukan kembali dokumen yang diinginkan. Representasi dokumen dilakukan dengan menghilangkan kata-kata yang tidak penting , kemudian menghitung bobot tiap term kemudian dimasukkan dalam tabel index sebagai representasi dokumen.

A R T I C L E I N F O

Article History:
Received
Revised
Accepted
Available online

Keywords:
Information Retrieval,
Rumah Sakit,
Vector Space Model

I. PENDAHULUAN

Information retrieval (IR) adalah penemuan bahan seperti dokumen yang bersifat terstruktur yang memenuhi kebutuhan informasi dari dalam koleksi besar yang tersimpan di dalam komputer (Manning, Raghavan, & Schütze, 2008). Banyaknya data yang diterima dari pengunjung yang memberikan penilaian pada pelayanan dan fasilitas yang diberikan dari pihak Rumaha Sakit Umum Haji Medan membuat pencarian atau pemilihan informasi ini tidak mungkin dilakukan secara manual karena kumpulan informasi yang sangat banyak, dan beragam. Dibutuhkan suatu sistem otomatis untuk membantu admin dalam menemukan informasi yang dibutuhkan. Untuk dapat melakukan pencarian berdasar substansi yang paling mirip, terdapat teknologi yang disebut information Text Retrieval. Information text retrieval adalah salah satu metode

yang digunakan untuk menyimpan data dengan cara memprosesnya (menghilangkan stop word) dan menyimpan tiap kata beserta informasi dari kata tersebut (letak kata, jumlah bobot, dll). Information retrieval berfokus pada proses yang terlibat di dalam representasi, media penyimpanan, mencari dan menemukan informasi yang relevan dari informasi yang diinginkan oleh admin. Sistem secara otomatis akan melakukan indexing secara offline dan temu kembali (retrieval) secara real time. Proses retrieval dimulai dengan mengambil query dari pengguna, menerapkan stop word removal sehingga dihasilkan keyword yang compaq tetapi dapat mewakili query tersebut, kemudian sistem menghitung kemiripan antara keyword dengan daftar dokumen yang diwakili oleh term-term di dalam index. Dokumen akan ditampilkan diurutkan berdasarkan dokumen yang paling mirip.

II. METODE PENELITIAN

2.1 Information Retrieval

Menurut (Brata & Hetami 2019) *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 (Brata & Hetami 2019) menjelaskan bahwa information retrieval terbagi dari beberapa bagian yang dijabarkan sebagai berikut:

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

2.2 Metode Pengolahan Data

Metode pengolahan data yang akan dilakukan pada penelitian ini terbagi menjadi dua, yaitu pengolahan data offline dan real time. Data yang diproses secara offline adalah data diberikan pengunjung Rumah Sakit Umum Haji Medan melalui kuesioner yang kemudian dimasukkan dalam database secara manual, proses ini disebut dengan indexing. sedangkan pengolahan data Query dilakukan secara real time

2.2.1. Pengolahan Data *Offline*

Adapun penjelasan mengenai pengolahan data secara *Offline* dapat dilihat sebagai berikut:

 I S S N : 2774-7751

a. Data yang terkumpul akan dilakukan preprocessing. Preprocessing meliputi penghilangan kata yang dianggap tidak penting (*stopword*) dan dilakukan *stemming*, yaitu mengubah kata ke bentuk dasarnya dengan cara menghilangkan imbuhan awal maupun imbuhan akhir. Dari proses ini akan dihasilkan daftar kata atau term yang lebih compaq tetapi tetap mewakili dokumen yang sedang diproses
b. Setelah dilakukan preprocessing, maka langkah selanjutnya adalah mengambil tiap kata/term dan menghitung jumlah kemunculannya pada dokumen tertentu.
c. Dilakukan pembobotan kata menggunakan rumus Tf/Idf. (1) Dimana Wij adalah bobot term j pada dokumen i Tfij adalah frekuensi term j pada dokumen i N adalah jumlah total dokumen yang dikoleksi Dfj adalah jumlah dokumen yang mengandung term. Tahapan indexing dilakukan untuk menyimpan tiap kata/term ke dalam database dengan atribut jumlah kemunculan dan bobot tiap term

2.2.2. Pengolahan Data *Realtime*

Query yang dimasukkan oleh user juga akan diolah melalui beberapa proses yaitu:

a. Melakukan preprosesing terhadap *query* yang dimasukkan user yaitu menghilangkan *stopword*.
b. Setelah dilakukan preprocessing, maka langkah selanjutnya adalah mengambil tiap kata/*term* dan menghitung jumlah kemunculannya pada dokumen tertentu.
c. Dilakukan pembobotan kata menggunakan rumus Tf/Idf.
d. Tahapan indexing dilakukan untuk menyimpan tiap kata/term ke dalam database beserta bobot tiap term.

Hal ini dilakukan dengan tujuan supaya query dengan kata yang sama tidak perlu dilakukan perhitungan lagi. Setelah data selesai diolah secara offline dan realtime, maka akan dilakukan perhitungan similaritas (kemiripan) antara query permintaan user dengan dokumen yang tersimpan dalam database. Perhitungan dilakukan menggunakan vector space model. Kemudian hasilnya akan ditampilkan beberapa dokumen yang relevan dengan query secara urut berdasarkan kemiripan

2.3. Pembobotan TF-IDF

Baeza-Yates dan Ribeiro-Neto menyebutkan bahwa pembobotan TF-IDF terdiri dari dua faktor, yaitu:

1. TF (*termfrequency*)

TF adalah frekuensi kemunculan suatu istilah fi di dalam sebuah dokumen djdibandingkan dengan frekuensi istilah fl yang sering muncul pada dokumen itu. Jika dimasukkan dalam rumus matematika didapatkan:

$$f_{i,j} = \frac{freq_{i,j}}{max_i\, freq_{i,j}}$$

Gambar. 2.1 Rumus Term Frequency

 I S S N : 2774-7751

2. IDF (*inverse document frequency*)
IDFadalah frekuensi kemunculan suatu istilah fi di dalam seluruh dokumen. Penggunaan faktor IDF didasarkan pada istilah yang muncul pada setiap dokumen tidak memberikan suatu ciri khusus untuk menentukan dokumenyang relevan dari yang tidak relevan. Jika jumlah seluruh dokumen di dalam sistem dinyatakan dengan nilai N dan jumlah dokumen yang memiliki istilah fi tersebut dinyatakan dengan ni, maka nilai IDFi-nya dapat dinyatakan dengan:

$$idf_i = log \frac{N}{n_i}$$

Gambar 2. *Rumus inverse document frequency*

IDF= *inverse document frequency*
N = jumlah kalimat yang berisi term(t)
ni= jumlah kemunculan kata (term) terhadap dj

Faktor pembobotan untuk tiap kata dalam dokumen didefinisikan sebagai kombinasi termfrequency dan inverse document frequency. Dari dua faktor tersebut maka pembobotan TF-IDF dapat dinyatakan dengan:

$$w_{i,j} = f_{i,j} \text{ x } log \frac{N}{n_i}$$

Gambar 3. Rumus TF-IDF

$$w_{ij} = tf_{ij} \text{ x } idf_j$$

Gambar 4. Rumus pembobotan TF-IDF

Keterangan:
W_{ij} = nilai bobot kata ke j dari dokumen i
Tf_{ij} = termfrequency,yakni jumlah kemunculan kata tj dalam dokumen Di
DF_j = document frequncy, yakni jumlah dokumen yang mengandung tj
IDF_j = log *d* dengan d adalah jumlah semua dokumen dalam koleksi. IDFj adalah inverse document frequency (Anistyasari dkk, 2012).

Pada Metode ini pembobotan kata dalam sebuah dokumen dilakukan dengan mengalikan nilai TF dan IDF. Pembobotan diperoleh berdasarkan jumlah kemunculan term dalam kalimat (TF) dan jumlah kemunculan term pada seluruh kalimat dalam dokumen (IDF). Bobot suatu istilah semakin besar jika istilah tersebut sering muncul dalam suatu dokumen dan semakin kecil jika istilah tersebut muncul dalam banyak dokumen (Fatkhul, 2019).

Kemudian baru melakukan proses pengurutan (sorting) nilai kumulatif dari W untuk setiap kalimat. Tiga kalimat dengan nilai W terbesar dijadikan sebagai hasil dari ringkasan atau sebagai output dari peringkasan teks otomatis (Sarno, dkk, 2012).

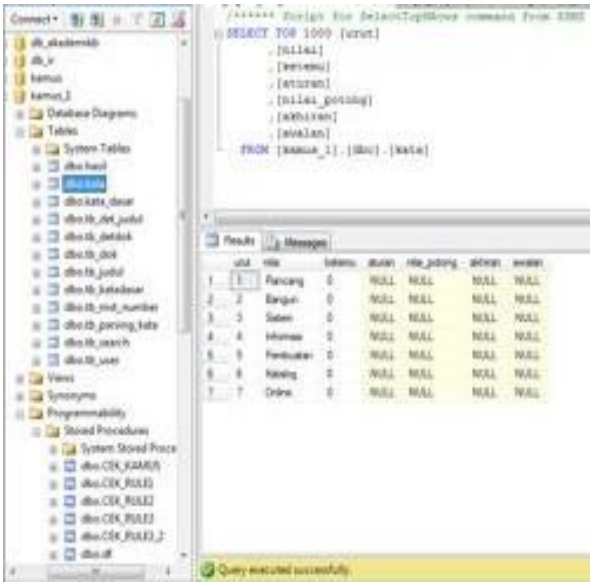
 ISSN : 2774-7751

III. HASIL DAN PEMBAHASAN

3.1. Implementasi Antar Muka

Berikut merupakan tampilan dari Aplikasi *Information Retrieval*. Terdapat menu Home yang berisi kolom pencarian, Arsip untuk melihat koleksi dokumen

Gambar 5.Tampilan A plikasi IR

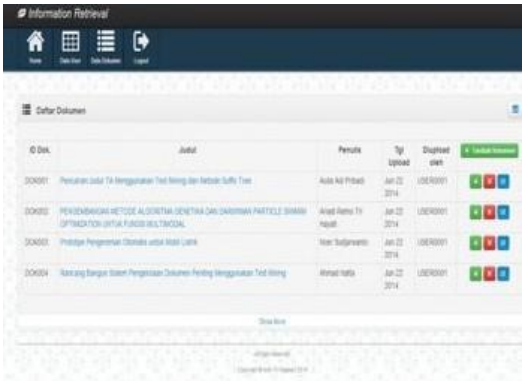


Gambar 6. Tampilan Data Dokumen

3.2. Proses *Teoknisi*

Hasil implementasi *information retrieval* saat proses *preprocessing* dokumen dan pembobotan TF-IDF adalah me-retrieve dokumen atau menemukan kembali dokumen yang diinginkan. Berikut merupakan pengujian tokenisasi dan *filtering* dengan memberikan inputan kalimat yang mengandung tanda baca dan kata yang termasuk *stoplist*.

Teks Awal : "perlunya peningkatan kebersihan dan pelayanan."

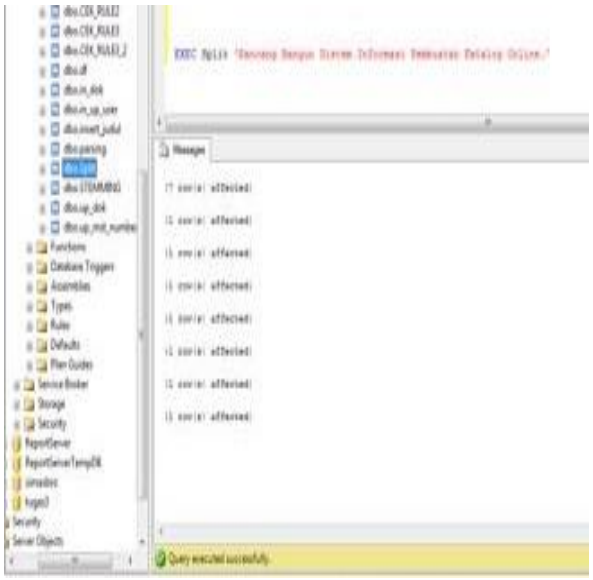


Gambar 7. Hasil Pengujian Tokenisasi

3.3. Proses *Filtering*

Adapun proses dari filtering dapat dilihat pada gambari dibawah ini: Teks Awal : "perlunya peningkatan kebersihan dan pelayanan."

 I S S N : 2774-7751



Gambar 8. Hasil Pengujian *Filtering*

IV. Kesimpulan

Dari perancangan dan implementasi yang telah dilakukan, maka dapat dibuat kesimpulan sebagai berikut :

1. Representasi dokumen dilakukan dengan menerapkan preprocessing, yaitu menghilangkan kata-kata yang tidak penting (*stopword*), kemudian dilakukan indexing, yaitu menghitung bobot (Tf/Idf) tiap term kemudian dimasukkan dalam tabel index sebagai representasi dokumen. *Query* yang dimasukkan oleh user juga akan diproses dengan cara yang sama (direpresentasikan dahulu) sebelum di hitung tingkat kemiripannya.
2. Proses perhitungan kemiripan dokumen dilakukan dengan mengetikkan *keyword, keyword* bisa terdiri dari satu kata atau lebih. *Keyword* yang dimasukkan juga akan diproses sama seperti dokumen. Kemudian dihitung tingkat kemiripan keyword dengan abstrak yang sesuai menggunakan rumus *vector space model* (VSM).

V. Daftar Pustaka

Amin, Fatkhul. 2011: Implementasi Search Engine (Mesin Pencari) Menggunakan Metode Vector Space Model, Vol. V No. 1, Hal 45- 48, Dinamika Teknik.

Baeza, Yates and Ribeiro, Neto. 1999. Modern Information retrieval. Harlow. Addison-Wesley. [2] Bunafit, Nugroho. 2008, Aplikasi Pemrograman Web Dinamis Dengan PHPdan MySQL. Gava Media. Yogyakarta Cholifah, 1978. “Aplikasi Informasi Retrieval Untuk Pembentukan Tesaurus Berbahasa Indonesia Secara Otomatis”.Scan vol II no.1, ISSN : 1978-0087.

Frakes, W.B. dan Baeza. R. 1992, Information retrieval Data Structure and Algorithms, New Jersey : Prentice Hall. [5] Grossman, D., 1992, IR Book, http://www.ir.iit.edu/~dagr/cs529/files/ir_book/ [7 Maret 2002]

Ingwersen, P, 1992, Information retrieval Interaction, London, Taylor Graham Publishing. http://www.db.dk/pi/iri [29 Agustus 2005]

Rijsbergen, C.J. van., 1979, Information retrieval, Second Edition. Butterworths, London. Jurnal SIMETRIS, Vol 8 No 1 April 2017 ISSN: 2252-4983 362

Salton, G., 1989, Automatic Text Processing : The transformation, Analysis, and Retrieval Information by Computer, Massachusetts, Addison-Wesley.

purwanti, endah,. 2015. "Klasifikasi Dokumen Temu Kembali dengan K-Nearest Neighbour". EISSN 2442-5168, Vol 1, No.1.

Salton, G. & Buckley, C., 1987, Term Weighting Approaches in Automatic Text Retrieval, Technical Report No. 87-881, Departement of Computer Science Cornell University Ithaca, New York.

Turney, P.D. Pantel, Patrick, 2010, "From Frequency To Meaning : Vector Space Model For Semantic" Journal of Artificial Intelligence Reseach, Vol 37, pp.141-188.

Welling, Luke and Thomson, Laura, 2001, PHP and MySQL Web Development.1st Edition. United States of America :Sams Publishing

Witten et all, 1999, Managing Gigabytes: Compressing and Indexing Document dan Images Second Edition, San Fransisco, Morgan Kaufmann Publishers.

Wisnu, dwija & Hetami, Anandhini. 2015," Perancangan Information retrieval (IR) untuk Pencarian Ide Pokok Teks Artikel Berbahasa Inggris Dengan Pembobotan Vector Space Model".Jurnal ilmiah Teknologi dan Informasi ASIA, Vol 9 No.1.