REPOSITOR, Vol. 2, No. 5, Mei 2020, Pp. 611-618
ISSN : 2714-7975
E-ISSN : 2716-1382

Rancang Bangun Tabloid Online Bestari dengan Fitur Pencarian berbasis Search Engine Teknologi menggunakan Metode Vector Space Model

Mentari Mas'ama Safitri[*1], Nur Hayatin[2], Yufis Azhar[3]
[1,2,3]Teknik Informatika/Universitas Muhammadiyah Malang
mentarims24@gmail.com[*1], noorhayatin@gmail.com[2], yufis.azhar@gmail.com[3]

Abstrak

Bestari merupakan lembaga pers mahasiswa yang merupakan media utama untuk menyuarakan dan mendokumentasikan berbagai kegiatan yang dilakukan sivitas akademika Universitas Muhammadiyah Malang. Bestari juga memliki tabloid online yang dapat diakses oleh mahasiswa, namun tabloid online Bestari Universitas Muhammadiyah Malang saat ini belum memiliki fitur pencarian, sehingga pengguna kesulitan untuk mendapatkan informasi sesuai dengan yang diinginkan. Berdasarkan masalah terse but pembangunan aplikasi Tabloid Online Bestari Universitas Muhammadiyah Malang dengan fitur pencarian berbasis search engine ini bertujuan untuk memberikan kemudahan kepada pengguna khususnya mahasiswa Universitas Muhammadiyah Malang dalam melakukan pencarian.Pada studi kasus ini metode Vector Space Model digunakan untuk memodelkan kumpulan berita dan keyword dari user dalam bentuk vektor yang telah di beri bobot dengan menggunakan metode pembobotan TF-IDF, kemudian akan di hitung kedekatan dari masing-masing dokumen dengan keyword dari user menggunakan cosine similarity.

Kata Kunci: *Bestari, Search Engine, VSM, TF-IDF, Cosine Similarity*

Abstract

Bestari is collage students' press agency that is the main media to show their opinion and document any activities that have already done by academic community of UMM. Bestari also has online tabloid that can be accessed by college students, but Bestari tabloid UMM does not currently have a search feature so the user difficult to get information needed. According to this problem, the developing of Tabloid Online Bestari UMM application that completed search feature with search engine based is purposed to give solution to the users especially for colledge students of UMM. This study used Vector Space Model (VSM) method to collect the news and keyword from the users in the form of vector that has been given scale used TF-IDF method, then the correlation of keyword with the document will be counted by using cosine similarity.

Keywords: *Bestari, Search Engine, VSM, TF-IDF, Cosine Similarity*

1. Pendahuluan

Penerapan teknologi digital dan jaringan komputer telah menyebabkan terjadinya "ledakan" informasi yang berkembang eksponensial. Hal ini menyebabkan Sistem temu kembali informasi **(***information retrieval=IR*) mengalami kesulitan [1]. *Information Retrieval* (IR) merupakan bagian dari *computer science* yang berhubungan dengan pengambilan informasi dari dokumen-dokumen yang didasarkan pada isi dan konteks dari dokumen-dokumen itu sendiri. Informasi yang diinginkan pengguna direpresentasikan dalam bentuk *query* dan mengandung satu atau lebih *term* yang akan digunakan dalam pencarian.

Search engine atau mesin pencari merupakan teknik dari temu-kembali dalam menemukan dokumen dan sekaligus mengeksekusi algoritma peringkat dalam menampilkan dokumen. Pengguna dapat mencari halaman *web* yang dibutuhkan melalui *search engine*. *Search engine* tidak lain sebuah mesin pencari yang ulet dan teliti, yang melakukan eksplorasi atas informasi-informasi yang di-*request* tanpa memandang kapan, di mana dan oleh siapa itu dilakukan. Mesin pencari menggunakan indeks (yang sudah dibuat dan disusun secara teratur) untuk mencari file setelah pengguna memasukkan kriteria pencarian. *Indexing* atau pengindeksan merupakan proses membangun basis data indeks dari koleksi dokumen. *Indexing* dilakukan terhadap dokumen sebelum pencarian dilakukan [2].

Berita dapat dikatakan sebagai kebutuhan pokok dalam diri seseorang. Karena manusia hidup pasti membutuhkan informasi. Penyebaran berita sekarang ini juga diiringi dengan kemajuan teknologi. Masyarakat bisa dengan mudah mengakses berita sebagai sumber informasi melalui media cetak dan elektronik. Perkembangan teknologi media elektronik mendukung penyebaran informasi melalui akses internet dari *gadget* oleh para pengguna. Mudahnya akses ini memerlukan keakuratan dalam hasil pencarian. Kecanggihan inilah yang digunakan oleh para instansi termasuk universitas untuk menyebarkan informasi.

Tabloid online Bestari Universitas Muhammadiyah Malang saat ini belum memiliki fitur pencarian, sehingga pengguna kesulitan untuk mendapatkan informasi sesuai dengan yang diinginkan. Berdasarkan masalah tersebut pembangunan aplikasi Tabloid Online Bestari Universitas Muhammadiyah Malang dengan fitur pencarian berbasis *search engine* teknologi ini bertujuan untuk memberikan kemudahan kepada pengguna khususnya mahasiswa Universitas Muhammadiyah Malang dalam melakukan pencarian berita. Dalam penelitian yang dilakukan oleh Heninggar Septiantri yang membandingkan metode LSA (*Latent Semantic Analysis*) dan VSM (*Vector Space Model*) mengenai sistem penilai jawaban esai otomatis Bahasa Indonesia, didapatkan dari hasil uji coba bahwa secara keseluruhan rata-rata korelasi nilai VSM-manusia lebih tinggi dari LSA-manusia [3]. Oleh karena itu metode yang akan digunakan untuk penelitian ini adalah metode *Vector Space Model.*

Vector Space Model adalah suatu metode untuk merepresentasikan sistem temu kembali ke dalam vektor dan memperhitungkan fungsi *similarity* dalam proses pencocokan beberapa vektor [3]. Untuk melakukan perhitungan fungsi *similarity* terlebih dahulu akan di lakukan perhitungan pembobotan dari masing-masing dokumen dan *keyword* menggunakan Metode Pembobotan TF-IDF. Metode TF-IDF (*Term Frequency-Inverse Document Frequency*) merupakan suatu cara untuk memberikan bobot hubungan suatu kata (*term*) terhadap dokumen [4]. Metode ini menggabungkan dua konsep untuk perhitungan bobot yaitu, frekuensi kemunculan sebuah kata didalam sebuah dokumen tertentu dan *inverse* frekuensi dokumen yang mengandung kata tersebut [5].

Pada studi kasus ini metode *Vector Space Model* (VSM) digunakan untuk memodelkan kumpulan berita dan *keyword* dari user dalam bentuk vektor yang telah di beri bobot dengan menggunakan metode pembobotan TF-IDF, kemudian akan di hitung kedekatan dari masing-masing dokumen dengan *keyword* dari user menggunakan *cosine similarity.* Untuk proses perangkingan hasil pencarian berita pada mesin pencari ini tidak hanya dilihat dari kedekatan antara *keyword* dan kumpulan berita, tetapi juga akan ditambahkan proses perhitungan *prosentase* untuk parameter tanggal posting dan jumlah *viewer* agar dapat menampilkan hasil dari pencarian berita yang sesuai dengan *input*-an dari *user*.

2. Metode Penelitian

2.1 TF-IDF (*Term Frequency – Inverse Term Frequency*)

Term Frequency adalah frekuensi dari kemunculan kata dari sebuah dokumen yang bersangkutan. Semakin besar jumlah kemunculan suatu term yang terdapat dalam dokumen, semakin besar pula bobotnya atau akan memberikan nilai kesesuaian yang semakin besar [6]. Rumus *Term Frequency* adalah Persamaan 1 [4].

$$w_{t,d} = \begin{cases} 1 + log_{10} tf_{t,d}, & if\ tf_{t,d} > 0 \\ 0, & if\ tf_{t,d} = 0 \end{cases} \quad (1)$$

Inverse Document Frequency merupakan perhitungan dari pendistribusian kata atau term secara luas pada koleksi dokumen yang bersangkutan. Semakin sedikit frekuensi dokumen yang mengandung kata atau *term* tertentu, maka nilai dari IDF semakin besar [5]. Rumus *Inverse Document Frequency* adalah Persamaan 2 [4].

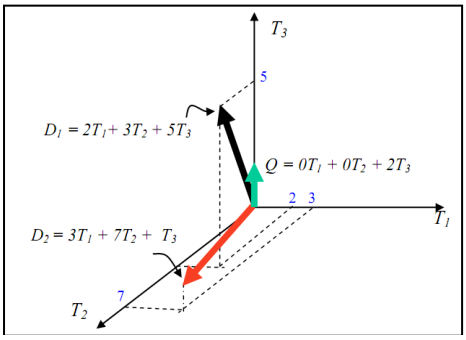
$$idf_t = log_{10}\left(\frac{N}{df_t}\right) \quad (2)$$

Sehingga bobot akhir suatu term adalah dengan mengalikan keduanya, yaitu tf x idf seperti pada Persamaan 3 [3].

$$W_{d,t} = tf_{d,t}\ x\ idf_t \quad (3)$$

2.2 *Vector Space Model* (VSM)

Metode Ruang Vektor pada Gambar 1 adalah suatu metode untuk merepresentasikan sistem temu kembali informasi. Relevansi dokumen dengan *query* dianggap sebagai ukuran kesamaan antara vektor dokumen dengan vektor *query*. Semakin sama antara vektor dokumen dengan vektor *query* maka dokumen dianggap semakin relevan dengan *query*.



Gambar 1. Representasi Dokumen dan Query pada Ruang Vektor (Sumber: Mandala dan Setiawan, 2002)

Perhitungan kesamaan antara vektor *query* dan vektor dokumen dilihat dari sudut yang terkecil. Sudut yang dibentuk oleh dua buah vektor dapat dihitung dengan melakukan perkalian dalam (inner product), sehingga rumus *Cosine Similarity* adalah Persamaan 4 [3].

$$similarity\left(\bar{d}_j, \bar{q}\right) = \frac{\bar{d}_j \,.\, \bar{q}}{|\bar{d}_j| \,.\, |\bar{q}|} = \frac{\sum_{i=1}^{t}(w_{ij} \,.\, w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \cdot \sum_{i=1}^{t} w_{iq}^2}} \tag{4}$$

Proses dilakukannya perangkingan dokumen dianggap sebagai proses pemilihan vektor dari suatu dokumen yang dekat dengan vektor dari *query* tersebut, kedekatan ini diindikasikan dari sudut yang dibentuk. Nilai *cosine* yang cenderung besar membuktikan bahwa dokumen cenderung sesuai *query*. Nilai *cosine* sama dengan 1 menyatakan bahwa dokumen sesuai dengan *query* [4].

2.3 *Recall* dan *Precision*

Recall digunakan untuk dokumen terpanggil yang relevan dengan *query* yang dimasukkan pengguna dalam sistem temu balik informasi. R*ecall* terhubung dengan kemampuan sistem untuk menemukan dokumen relevan. *Recall* adalah bagian dari proses temu balik informasi yang bisa digunakan sebagai alat ukur untuk tingkat efektivitas suatu sistem. *Recall* berhubungan dengan kemampuan sistem memanggil dokumen relevan, sedangkan *precision* berkaitan dengan kemampuan sistem untuk tidak memanggil dokumen tidak relevan. Sebenarnya *recall* sulit diukur karena jumlah semua dokumen relevan yang terdapat dalam database sangat besar. Oleh karena itu *precision* yang menjadi suatu ukuran yang digunakan untuk memberika nilai keefektifan suatu sistem. *Precison* adalah jumlah frekuensi dokumen yang relevan dari jumlah frekuensi dokumen yang ditemukan oleh sistem. Presisi juga merupakan cara mengukur tingkat efektivitas sistem temu balik informasi [7].

Recall menemukan seluruh dokumen yang relevan dalam koleksi. Recall dapat dihitung dengan Persamaan 5 [8].

$$Recall = \frac{Jumlah\ dokumen\ relevan\ yg\ terpanggil}{Jumlah\ dokumen\ relevan\ yg\ ada\ di\ dalam\ database}\ x\ 100 \tag{5}$$

Nilai recall tertinggi adalah 1, yang berarti seluruh dokumen dalam koleksi berhasil ditemukan.

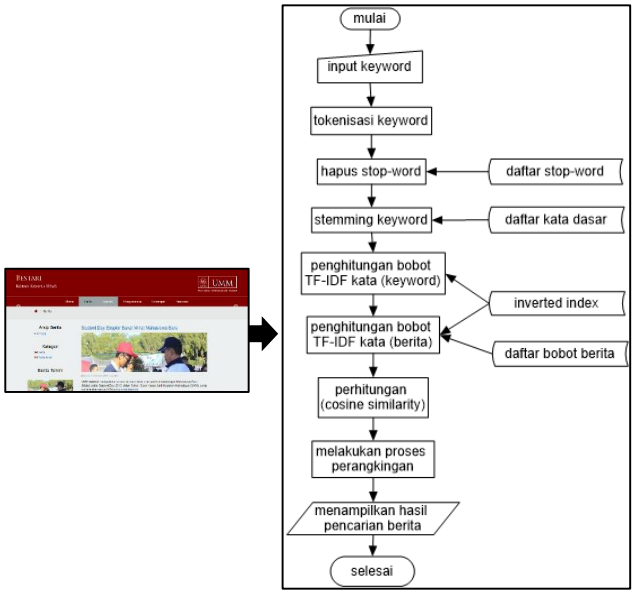
Precision hanya menemukan dokumen yang relevan saja dalam koleksi. Precision dapat dihitung dengan Persamaan 6 [8].

$$Precision = \frac{Jumlah\ dokumen\ relevan\ yg\ terpanggil}{Jumlah\ dokumen\ relevan\ yg\ terpanggil\ dalam\ pencarian}\ x\ 100 \quad (6)$$

Nilai precision tertinggi adalah 1, yang berarti seluruh dokumen yang ditemukan adalah relevan.

2.4 Gambaran umum aplikasi

Pada tahapan ini dilakukan analisa desain aplikasi search engine yang akan dibangun. Analisis sistem akan digambarkan dengan diagram alur, agar alur proses sistem dapat lebih mudah dipahami Gambar 2 merupakan rancangan proses sistem "Tabloid Online Bestari dengan Fitur Pencarian berbasis *Search Engine* Teknologi".



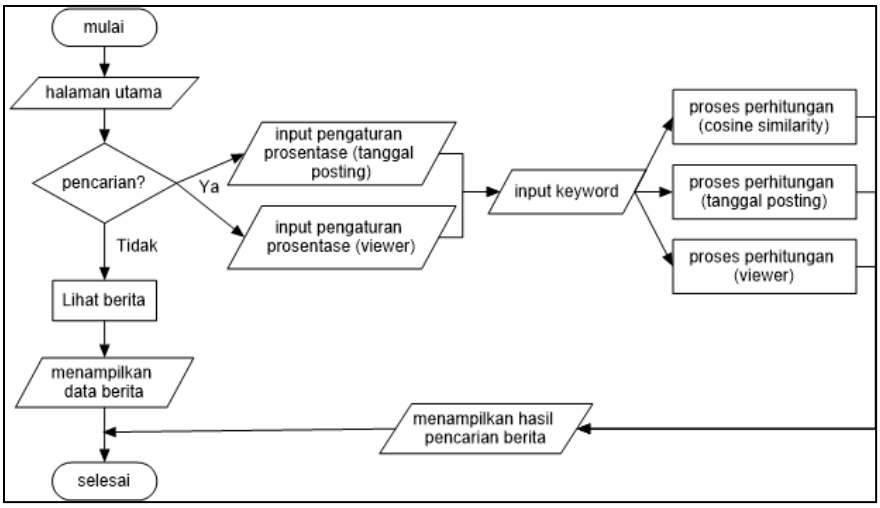
Gambar 2. Perancangan Alur Proses Sistem "Tabloid Online Bestari dengan Fitur Pencarian berbasis Search Engine Teknologi"

Deskripsi keterangan untuk Gambar 2 diatas adalah sebagai berikut, user memasukkan keyword untuk melakukan pencarian, setelah keyword dimasukkan dilakukan tahap pemotongan teks input (keyword) berdasarkan tiap kata yang menyusunnya yang disebut dengan tokenisasi. Kemudian dilakukan penghapusan kata umum dan dianggap tidak memiliki makna (*stopword*), selanjutnya dilakukan proses stemming yaitu untuk menemukan kata dasar dari sebuah kata. Setelah melakukan semua tahapan di atas, dilanjutkan ke tahapan berikutnya yaitu perhitungan pembobotan suatu kata dari keyword yang dimasukkan oleh user dan kata dari dokumen berita yang didapatkan dari tabloid bestari menggunakan metode pembobotan kata TF-IDF (*Term Frequency-Inverse Document Frequency.* Setelah proses perhitungan, dilakukan pengurutan dokumen sesuai dengan hasil perhitungan *cosine similary*, dimana semakin besar hasilnya maka jarak antara dokumen terhadap *keyword* tersebut semakin dekat, selain itu akan ditambahkan proses perhitungan *prosentase* untuk parameter tanggal posting dan jumlah *viewer* agar dapat menampilkan hasil dari pencarian berita yang sesuai dengan *input*-an dari *user*. Jadi hasil dari dokumen terurut yang akan ditampilkan oleh sistem tidak hanya dilihat dari kemiripan antar dokumen terhadap *keyword*, tetapi juga akan disesuaikan dengan tanggal *posting* dan jumlah *viewer*.

2.5 Diagram Alir Halaman User

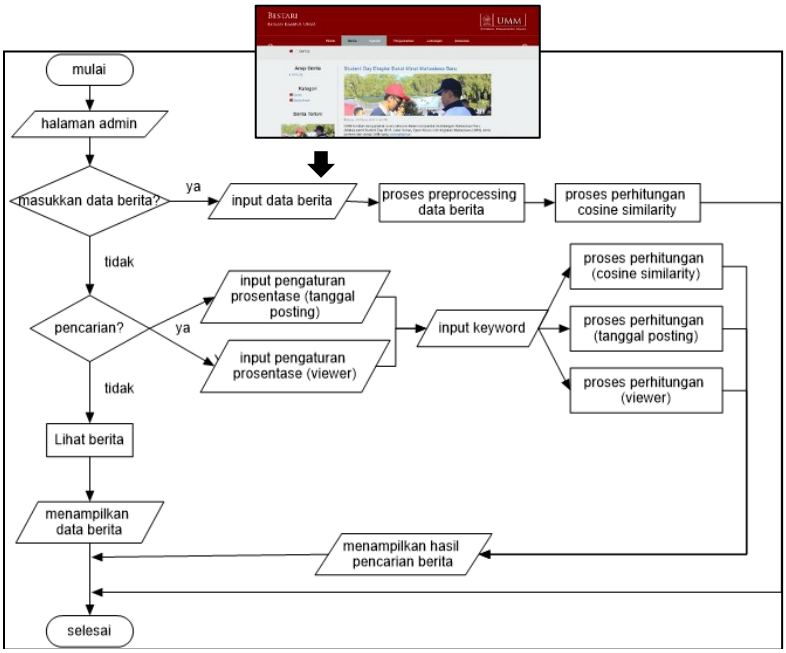
Deskripsi keterangan untuk Gambar 3 diatas adalah sebagai berikut, user akan masuk ke halaman user. Pada halaman user terdapat 2 fitur, yaitu pencarian dan berita. Jika user memilih fitu pencarian, maka user akan melakukan input prosentase untuk perhitungan tanggal dan jumlah view, sebelumnya untuk prosentase perhitungan cosine similarity sudah diset 100%, jadi jika user tidak melakukan input prosentase tanggal dan viewer, otomatis sistem akan menganggap prosentase tanggal dan viewer adalah 0%. Setelah itu input keyword untuk

melakukan pencarian, kemudian dilakukan proses perhitungan cosine similarity, tanggal posting dan jumlah viewer (jika prosentase sudah terisi). Selanjutnya sistem akan menampilkan hasil dari pencarian berita. Jika tidak, maka user bisa memilih fitur lihat berita yang dapat menampilkan kumpulan data berita yang terdapat pada sistem ini.



Gambar 3. Diagram Alir Halaman User

2.6 Diagram Alir Halaman Admin



Gambar 4. Diagram Alir Halaman Admin

Deskripsi keterangan untuk Gambar 4 diatas adalah sebagai berikut, setelah melakukan login sebagai admin, kemudian akan masuk ke halaman admin. Pada halaman admin ada beberapa fitur, yang pertama adalah fitur untuk melakukan input data berita. Jika admin memilih fitur tersebut, selanjutnya admin akan memasukkan data berita (data berita yang akan dimasukkan diambil dari data berita pada koran bestari UMM), setelah data berita dimasukkan akan dilakukan proses preprocessing pada data berita, kemudian yang terakhir akan dilakukan perhitungan bobot pada masing-masing kata yang terdapat pada data berita dan dilakukan perhitungan cosine similarity-nya. Jika fitur untuk melakukan input data berita tidak dipilih, maka akan dilanjutkan pada fitur selanjutnya, yaitu fitur pencarian. Jika iya admin input prosentase untuk perhitungan tanggal dan jumlah view, sebelumnya untuk prosentase perhitungan cosine

similarity sudah diset 100%, jadi jika tidak melakukan input prosentase tanggal dan viewer, maka otomatis sistem akan menganggap prosentase tanggal dan viewer adalah 0%. Setelah itu input keyword untuk melakukan pencarian, kemudian dilakukan proses perhitungan cosine similarity, tanggal posting dan jumlah viewer (jika prosentase sudah terisi). Selanjutnya sistem akan menampilkan hasil dari pencarian berita. Jika kedua fitur diatas tidak dipilih, maka admin bisa memilih fitur lihat berita yang dapat menampilkan kumpulan data berita yang terdapat pada sistem ini.

3. Hasil Penelitian dan Pembahasan

Pengujian dilakukan untuk mengukur kemampuan atau tingkat kinerja aplikasi dilakukan dengan 10 kata kunci yang telah ditetapkan, dimana hasil dari pencarian yang menggunakan metode *cosine similarity* dengan prosentase perhitungan sebesar 100% dari masing-masig *keyword* akan dievaluasi nilai dari *precision* dan *recall* untuk mengetahui tingkat kinerja dari aplikasi. Tabel 1 merupakan hasil dari perhitungan *precision* dan *recall*.

Tabel 1. Hasil Perhitungan Precision dan Recall

No.	*Keyword*	*Precision* (%)	*Recall* (%)
1	Teknik Informatika	$\frac{4}{27}x100\% = 14{,}81$	$\frac{4}{4}x100\% = 100$
2	Asupan Gizi	$\frac{2}{11}x100\% = 18{,}19$	$\frac{2}{2}x100\% = 100$
3	Kapal Militer	$\frac{1}{11}x100\% = 9{,}090$	$\frac{1}{1}x100\% = 100$
4	Mutu Pengajar	$\frac{10}{120}x100\% = 8{,}334$	$\frac{10}{10}x100\% = 100$
5	Negeri Indonesia	$\frac{42}{176}x100\% = 23{,}863$	$\frac{42}{42}x100\% = 100$
6	Mahasiswa Wirausaha	$\frac{14}{183}x100\% = 7{,}650$	$\frac{14}{14}x100\% = 100$
7	Luar Negeri	$\frac{4}{69}x100\% = 5{,}797$	$\frac{4}{4}x100\% = 100$
8	Hukuman Mati	$\frac{13}{78}x100\% = 16{,}67$	$\frac{13}{13}x100\% = 100$
9	Bangsa Indonesia	$\frac{43}{164}x100\% = 26{,}219$	$\frac{43}{43}x100\% = 100$
10	Potensi Diri	$\frac{8}{82}x100\% = 9{,}756$	$\frac{8}{8}x100\% = 100$
	Jumlah	140,379	1000
	Rata-rata	14,03	100

Dari hasil uji perhitungan yang dipaparkan pada Tabel 1 dapat dilihat rata-rata dari nilai *precision* dan *recall*. Nilai rata-rata *precision* dari aplikasi berdasarkan uji coba terhadap *keyword* adalah 0,1403, atau bisa dikatakan aplikasi pencarian tabloid online bestari ini memiliki tingkat keakuratan dengan persentase 14,03 %. Nilai precision yang terlalu kecil dikarenakan sistem menggunakan metode pencarian *cosine similarity*, pada metode ini sistem hanya bisa membandingkan tingkat kemiripan antara *keyword* dengan dokumen berita yang ada, tanpa melihat persamaan makna antara *term* yang ada pada *keyword* dengan *term* yang ada pada dokumen berita. Sehingga pada contoh pencarian pertama untuk *keyword* “Teknik Informatika”, jumlah dokumen relevan yang terpanggil dalam pencarian sebanyak 27 dari 300 data berita yang ada, sedangkan jumlah dokumen relevan yang terpanggil sesuai dengan *keyword* “Teknik Informatika”, sebanyak 4 data berita dari 27 dokumen relevan yang terpanggil. Untuk nilai *recall* dari pengujian ini dengan menggunakan metode *cosine similarity* menghasilkan nilai *recall* yang tinggi, karena nilai dari jumlah dokumen relevan yang terpanggil sama besar dengan jumlah dokumen relevan yang ada didalam database. Sehingga menghasilkan nilai *recall* sebesar 100%.

4. Kesimpulan

Berikut ini merupakan kesimpulan dari implementasi dan uji coba terhadap aplikasi pencarian berita pada tabloid *online* bestari yang telah dirancang. Dengan menggunakan metode

Vector Space Model aplikasi ini mampu melakukan pencarian terhadap dokumen berita yang ada pada tabloid bestari Universitas Muhammadiyah Malang. Berdasarkan hasil uji coba terhadap 10 *keyword* dimana nilai rata-rata tingkat keakuratan aplikasi ini terhadap hasil pencarian mencapai nilai persentase 14,03%. Sedangkan tingkat kemampuan aplikasi untuk menampilkan semua data berita yang relevan dengan *keyword* mencapai nilai persentase sebesar 100%. Aplikasi dapat memudahkan *user* dalam mencari informasi mengenai berita seputar kampus Muhammadiyah Malang dan berbagai kegiatan yang mahasiswa, serta tips dan saran untuk kesehatan

5. Saran

Untuk pengembangan lebih lanjut mengenai aplikasi pencarian pada tabloid *online* bestari Universitas Muhammadiyah Malang ini, diberikan saran-saran sebagai berikut Aplikasi ini tidak hanya bisa melakukan pencarian dokumen dalam bahasa Indonesia, tetapi juga bisa dilakukan pencarian dokumen dalam bahasa Inggris. Aplikasi ini dapat dikembangkan menggunakan metode lain yang bisa memberikan tingkat keakuratan hasil pencarian yang lebih tinggi dibandingkan dengan metode sebelumnya. Dan aplikasi ini dapat dikembangkan menggunakan metode lain yang bisa melakukan pencarian tidak hanya menggunakan tingkat kemiripan suatu dokumen, tetapi juga bisa melakukan pencarian berdasarkan persamaan makan.

Daftar Notasi

d : dokumen ke-d
t : kata ke-t dari kata kunci
$W_{d,t}$: bobot dokumen ke-d terhadap kata ke-t
tf : banyaknya kata yang dicari pada sebuah dokumen
idf : *Inverse Document Frequency*
D : total dokumen
df : banyak dokumen yang mengandung kata yang dicari
d_j : dokumen ke-j
q : kata kunci
t : term / kata
w_{ij} : pembobotan kata ke-i dari dokumen ke-j
w_{iq} : pembobotan kata ke-i dari kata kunci

Referensi

[1] Aminx, F. (2012). Sistem Temu Kembali Informasi dengan Metode *Vector Space Model*. *Jurnal Sistem Informasi Bisnis 02* , 78-83.
[2] Fitri, M. (2013). Perancangan Sistem Temu Balik Informasi dengan Metode Pembobotan Kombinasi TF-IDF untuk Pencarian Dokumen Berbahasa Indonesia.
[3] Darmawan, H. A., Wurijanto, T., & Masturi, A. (2010). Rancang Bangun Aplikasi Search Engine Tafsir Al-Qur'an Menggunakan Teknik Text Mining Dengan Algoritma VSM (Vector Space Model).
[4] Karmayasa, O., & Mahendra, I. B. (n.d.). Implementasi Vector Space Model dan Beberapa Notasi Metode Term Frequency Inverse Document Frequency (TF-IDF) pada Sistem Temu Kembali Informasi.
[5] Hamzah, A. (2009). Temu Kembali Informasi Berbasis Kluster untuk Sistem Temu Kembali Informasi Teks Bahasa Indonesia. *Jurnal Teknologi*.
[6] (2014, september 25). Retrieved from rahmadya: https:/rahmadya.com/2014/09/25/term-frequency-dan-invers-document-frequency-tf-idf/
[7] (2010, juni 9). Retrieved from repository.usu.ac.id: http://repository.usu.ac.id/bitstream/handle/123456789/17855/Chapter%20II.pdf
[8] Wibowo, A. (2011). Pengujian Kerelevanan Sistem Temu Kembali Informasi.

ISSN: 2714-7975; E-ISSN: 2716-1382