PENCARIAN LINK INFORMASI PADA APLIKASI ENSIKLOPEDIA BUAH DAN SAYURAN LOKAL DENGAN METODE *COSINE SIMILARITY*

Agus Salim[1], Wiwik Suharso[2], Hardian Oktavianto[3]
Jurusan Teknik Informatika Fakultas Teknik Universitas Muhammadiyah Jember
aguss6460@gmail.com, wiwiksuharso@unmuhjember.ac.id, hardian@unmuhjember.ac.id

ABSTRAK

Pada fiitur pencarian link informasi di aplikasi ensiklopedia buah dan sayuran lokal Jember. Semua user dapat mencari link informasi buah dan sayuran lokal dengan kata kunci secara bebas. Analisis sistem ensiklopedia pada pencarian link informasi ini menggunakan metode *cosine similarity*. Penelitian dimulai dari *stemming* untuk mencari kata dasar dan TF*IDF untuk mencari nilai bobot di setiap *term* yang akan menghasilkan nilai *cosine* di setiap dokumen. Analisis hasil pengujian rata-rata kinerja *precision* terbaik dari skenario 1, 2 dan 3 sebesar 87% dengan ambang batas 0,3 sehingga digunakan oleh pengguna sebagai default pencarian dengan *precision* tertinggi. Sedangkan hasil pengujian rata-rata kinerja *recall* terbaik dari skenario 1, 2 dan 3 sebesar 100% dengan ambang batas 0,1 sehingga digunakan oleh pengguna sebagai default pencarian dengan *recall* tertinggi. Akan tetapi nilai rata-rata terbaik dari kinerja *precision* dan *recall* sebesar 87% dan 94% dengan ambang batas 0,3 sebagai default pencarian pada aplikasi ensiklopedia buah dan sayuran lokal Jember**.**

Kata kunci : Buah dan Sayuran lokal Jember, Ensiklopedia, *Cosine similarity.*

ABSTRACT

*At the search feature information link in the application of encyclopedia of local fruits and vegetables Jember. All users can search for free local fruit and veg information links with keywords. Analysis of encyclopedic system on this information link search using cosine similarity method. Research starts from stemming to find the base word and TF * IDF to find the weight value in each term that will generate the cosine value in each document. The analysis of the best average performance precision test results from scenario 1, 2 and 3 is 87% with a threshold of 0.3 so that it is used by the user as the highest search default with precision. While the best average recall performance results from scenario 1, 2 and 3 are 100% with a threshold of 0.1 so it is used by the user as the default search with the highest recall. However, the best average value of precision and recall performance is 87% and 94% with a threshold of 0.3 as the search default on the local Jember fruit and vegetable encyclopedia application.*

Keywords : *Local Fruits and Vegetables Jember, Encyclopedia, Cosine Similarity.*

1 PENDAHULUAN

Ensiklopedia sudah di kenal oleh kalangan pelajar sebagai media untuk mendapatkan informasi tentang topik tertentu yang di inginkan. Kebanyakan produk Ensiklopedia di pasaran dalam bentuk buku, majalah, atlas dan kartu. Produk ensiklopedia fisik tersebut telah digunakan dalam proses pembelajaran siswa-siswi di sekolah. Akan tetapi produk ensiklopedia fisik memiliki keterbatasan dalam kemudahan akses dan kecepatan penyebaran informasi serta bersifat statis (Suharso, 2017). Sementara Ensiklopedia online seperti wikipedia memiliki kelemahan dalam manajemen penyuntingan, akurasi informasi, dan informasi yang bersifat umum atau tidak spesifik pada konten data lokal (Mauludin, 2012). Oleh karena itu, penelitian ini bertujuan untuk membangun aplikasi Ensiklopedia buah dan sayuran lokal Jember berbasis web. Harapannya kalangan pelajar dan

masyarakat dapat mencari berbagai informasi muatan lokal tentang buah dan sayuran di Kabupaten Jember secara mudah, cepat dan akurat sebagai referensi pembelajaran. Aplikasi Ensiklopedia berbasis web ini dapat diakses secara bersamaan, dimanapun, kapanpun selama terhubung dengan jaringan internet. Agar informasi hasil pencarian memiliki akurasi yang tinggi, maka penelitian ini mengunakan metode *Cosine Similarity*. Sarno (2008) menanyatakan bahwa Metode *Cosine Similarity* tersebut dilakukan dengan mempertimbangkan frekuensi suatu kata dalam suatu dokumen (*term frequency*), dan penyebaran suatu kata pada sekumpulan dokumen (*Inverse Document Frequency*) serta kemiripan antar dokumen dengan metode *Cosine Similarity* dinyatakan dalam nilai bobot dari TF*IDF. Perhitungan kemiripan antara *Query* dengan *Document* dalam aplikasi Ensiklopedia dengan metode *Cosine Similarity* akan menghasilkan informasi buah dan sayuran lokal Jember secara akurat

1.1 .Rumusan Masalah Penelitian

Dalam penelitian ini rumusan masalahnya adalah bagaimana caranya membangun perangkat lunak ensiklopedia buah dan sayuran lokal Jember berbasis web dan pemanfaatan metode komputasi *Cosine Similarity* dalam menghasilkan informasi pencarian.

1.2 Batasan Masalah Penelitian

Batasan-batasan dalam penelitian ini sebagai berikut.

1. Bahasa pemrograman yang digunakan adalah *PHP Script*, dan DBMS yang digunakan adalah *Database MariaDB*.
2. Metode pencarian informasi berdasarkan *Query* yang akan dipertimbangkan dengan hasil komputasi *Cosine Similarity*.
3. Dataset penelitian menggunakan data buah dan sayuran lokal Jember dari penelitian ANIK ANDRIANI (1310211030) dan KUTSIATUL HIDAYAH (1310211006).

1.3 Tujuan Penelitian

Adapun tujuan penelitian sebagai berikut :

1. Membangun perangkat lunak ensiklopedia buah dan sayuran lokal Jember berbasis web sebagai sumber pembelajaran.
2. Memanfaatkan metode *Cosine Similarity* dalam proses komputasi untuk menghasilkan informasi yang akurat.

1.4 Manfaat Penelitian

Adapun manfaat penelitian sebagai berikut :

1. Kalangan pelajar dapat menggunakan produk ensiklopedia untuk membantu dalam mendapatkan informasi tentang topik buah dan sayuran lokal jember dalam beragam jenis seperti teks dan gambar secara mudah dan cepat.
2. Sistem ensiklopedia dapat memberikan link informasi secara akurat berdasarkan kata kunci dengan metode C*osine Similarity.*

2. TINJAUAN PUSTAKA

2.1 Data Buah dan Sayuran

Sawitri (2017) menyatakan bahwa Buah dan Sayuran lokal Jember merupakan merupakan bahan pangan utama dalam kehidupan sehari-hari. Jenis buah dan sayuran ini memiliki kurang lebih 107 macam jenis yang keanekaragamanya sangat bervariasi. Keanekaragaman warna pada buah bukanlah sekedar pembeda jenis antar buah yang satu dengan yang lainnya. Data Buah dan Sayur ini sudah banyak di sajikan berupa Buku Bacaan, Majalah, Atlas dan lain-lain.

2.2 Web Ensiklopedia

Mauludin (2012) menyatakan bahwa Web Ensiklopedia yang berbasis Teknologi Informasi (TI) merupakan tempat untuk mecari informasi sesuai kebutuhan sehari-hari, sehingga pengunjung web tersebut dapat mengendalikan situs tersebut sesuai fitur-fitur yang di sediakan. Beberapa situs website yang meyediakan layanan yang mengandung Ensiklopedia sebagai berikut :

http://www.wikipedia.org wikipedia.org merupakan situs yang mengandung layanan Ensiklopedia,**.** Ensiklopedia bebas berbahasa Indonesia dan di bangun oleh para sukarelawan. serta gratis. Salah satu keunggulannya adalah di

sediakan fasilitas pencarian, seperti situs search engine.

Kelemahan :

1. Karena platform itu memungkinkan siapun dapat terlibat dalam penulisan/pengeditan.
2. Kurangnya kesimpulan yang mendetail.
3. Kepengarangan/sumber informasi kurang di kenal.

2.3 Cosine Similarity

Triana (2016) menyatakan bahwa Metode *Cosine similarit***y** merupakan metode yang digunakan untuk menghitung tingkat kesamaan (*similarity*) antar dua buah objek. *Query* dan *Document* di proses dengan memanfaatkan *Text Mining* yang sebagai pra-proses.

Text mining didefinisikan sebagai proses penemuan kembali relasi dan fakta yang terkubur didalam teks dan tidak harus baru. Dalam penelitian ini *text mining* meliputi *tokenizing, stoplist/wordlis/Filt, stemming*. Dalam penelitian ketiga operasi *text mining* tersebut sebagai tahap persiapan untuk menyaring dan mengurangi jumlah informasi atau term-term yang tidak memiliki relevansi dalam pengukuran derajat kemiripan suatu dokumen. Proses selanjutnya adalah pembobotan kata dan pengukuran kemiripan dokumen.

2.3.1 *Tokenizing*

Tokenizing adalah proses pemecahan dokumen atau paragraf atau kalimat menjadi daftar kata atau *token* yang berdiri sendiri. Fungsi *token*tersebut akan melakukan pengecekan terhadap jarak antar kata (spasi, tabulasi, enter) untuk membuat daftar kata pada semua kata yang terdapat dalam dokumen.

2.3.2 Stoplist*/Wordlist/Filtering*

Stoplist/Wrodlist/Filtering adalah daftar kata yang tidak relevan dalam mengidentifikasi isi suatu dokumen seperti kata sambung, kata depan, kata ganti, simbol dan tanda baca. *Wordlist* adalah daftar kata yang relevan sebagai kata kunci sehingga akan digunakan dalam stemming. Stoplist/Wordlist adalah proses pembuangan atau penyaringan terhadap *stoplist* sehingga dihasilkan *wordlist*.

2.3.3 *Stemming*

Stemming bertujuan mengubah atau mengembalikan kata menjadi bentuk dasarnya dengan menghilangkan imbuhan-imbuhan pada kata dalam dokumen. Pembentukan kata dasar ini menggunakan algoritma porter dari Dr. Martin Porter yang disesuaikan atau dikembangkan dalam Bahasa Indonesia.

Sistem pencarian informasi secara otomatis dapat dilakukan dengan membandingkan *content identifier* berupa kata *(term*) yang terdapat pada teks (*document*) dan informasi yang diminta oleh user (*user information queries*). Dokumen dapat berupa dokumen, paragraf atau kalimat *D* dinyatakan dalam *term vectors*.

$$D = (t_i, t_j, ...t_p) \quad (1)$$

Dimana setiap t_k mengidentifikasikan term yang terdapat pada dokumen*D*. Demikian juga pada *query**Q* direpresentasikan dalam *term vectors*.

$$Q = (q_a, q_b, ...q_r) \quad (2)$$

Dimana setiap q_k mengidentifikasi term yang terdapat pada query *Q*. Sehingga bobot (*weight*) pada setiap term untuk membedakan term yang terdapat dalam dokumen *D* dan query*Q* dituliskan sebagai berikut.

$$D = (t_0, w_{d0}; t_1, w_{d1}; ...; t_t, w_{dt}) \quad (3)$$

$$Q = (q_0, w_{q0}; q_1, w_{q1}; ...; q_t, q_a, w_{qt}) \quad (4)$$

Dimana W_{dk} merupakan bobot dari term t_k dalam dokumen*D*, dan W_{qk} merupakan bobot term t_k dalam dokumen*Q*.

Metode TF-IDF (*Vecktor Space Model*) adalah cara untuk memberikan bobot hubungan suatu kata (term) terhadap dokumen. Proses pembobotan ini membentuk bobot $\omega_D(t_i)$ melibatkan tiga tahapan yaitu TF, IDF, dan TF*IDF dengan persamaan :

$$TF = (t_i, D) \quad (5)$$

$$IDF = \log \frac{N}{df(t_i, D)} + 1 \quad (6)$$

$$\omega_D(t_i) = \frac{tf(t_i, D) x \log \frac{N}{df(t_i)} + 1}{\sqrt{\sum_{t_i} (tf(t_i, D) x \log \frac{N}{df(t_i)} + 1)^2}} \quad (7)$$

Penilaian tingkat kemiripan *query-document* bisa didapatkan dengan membandingkan antara kedua vektor yang sesuai dengan persamaan 8.

$$Cos(Q, D) = \sum_{r=1}^{M} \omega_Q(t_i) x \omega_D(t_i) \quad (9)$$

2.4 *Recall* dan *Precison*

Raharjo (2013) menyatakan bahwa dalam "dunia" pengenalan pola (*pattern recognition*) dan temu kembali informasi (*information retrieval*), *precision* dan *recall* adalah dua perhitungan yang banyak digunakan untuk mengukur kinerja dari sistem / metode yang digunakan.

Recall adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi.

Precision adalah tingkat *ketepatan* antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem.

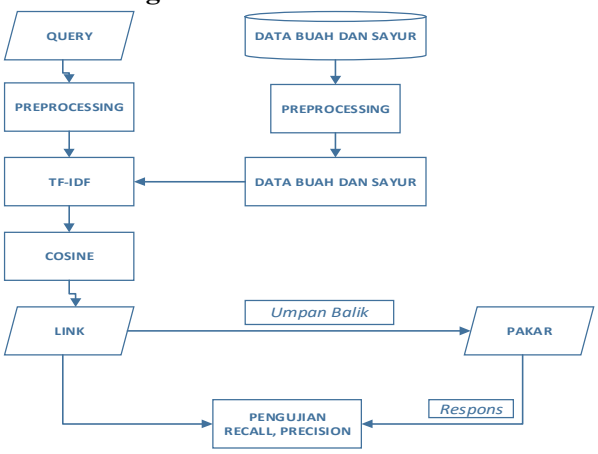
Secara umum *precision* dan *recall* dapat dirumuskan sebagai berikut:

		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*

Gambar 2.4 Recall dan Precision

3 METODELOGI PENELITIAN

3.1 Rancangan Sistem



Gambar 3.1 Rancangan Sistem

3.2 Skenario Pengujian

Skenario pengujian dilakukan dengan membuat 10 test-case query pengguna yang berbeda baik jumlah kata dan jenis data yang diinginkan. Ambang batas (threshold) nilai similarity menggunakan rata-rata keseluruhan nilai cosine dari link-link yang dihasilkan oleh sistem. Hasil pencarian dari masing-masing test-case akan dilakukan perbandingan kinerja dengan precision dan recall.

Diberikan contoh data perhitungan dalam menghasilkan link-link informasi yang dibutuhkan berdasarkan query pengguna sebagai berikut.

3.2.1 Preprocessing Query dan Dokumen.

Q : "Buah Buahan angga Manalagi"

a. Tokenizing : "buah manga gadung"
b. Stoplist/Stopword/Filtering: "buah mangga manalagi"
c. Unique : "buah mangga manalagi"
d. Stemming : "buah manga gadung"

D1 : Nama : "MANGGA GADUNG"

Abstrak: "Mangga Gadung Mangga gadung merupakan buah lokal. Buah ini digolongkan sebagaitermasuk buah sejati tunggal berdaging. Kulit luar buah berwarna hijau dan daging buah berwarna oranye (Hermanto, 2013). Dari empat kecamatan yang dijadikan sampel, mangga gadung hanya ditemuakan di Kecamatan Kaliwates."

a. Tokenizing, Stoplist/Stopword/Filtering, Stemming

Nama : "manga gadung"

Abstrak : "mangga gadung mangga gadung rupa buah lokal buah golong sebagaitermasuk buah sejati tunggal daging kulit luar buah warna hijau daging buah warna oranye hermanto 2013 empat camat sampel mangga gadung tua di camat kaliwates"

D2 : Nama : "MANGGA MANALAGI"

Abstrak : "Mangga Manalagi Mangga manalagi merupakan buah lokal. Buah ini digolongkan sebagaibuah sejati tunggal berdaging. Kulit luar buah berwarna hijau dan daging buah berwarna oranye (Hermanto, 2013). Dari empat kecamatan yang dijadikan

sampel, mangga manalagi ditemuakan di Kecamatan Tanggul, Kaliwates, Ambulu dan Balung"

a. Tokenizing, Stoplist/Stopword/ Filtering, Stemming

Nama : "manga manalagi"

Abstrak : "mangga manalagi mangga manalagi rupa buah lokal buah golong ibuah sejati tunggal daging kulit luar buah warna hijau daging buah warna oranye hermanto 2013 empat camat yang jadi sampel mangga manalagi tua camat tanggul kaliwates ambulu balung"

D3 : Nama : "MELON ROCK"

Abstrak : "Melon rock Melon rock merupakan buah lokal. Buah ini digolongkan sebagai buah sejati tunggal berdaging. Kulit luar buah berwarna hijau ke abu-abuan dan daging buah berwarna oranye (Hermanto, 2013). Dari empat kecamatan yang dijadikan sampel, melon rock hanya ditemuakan di Kecamatan Kaliwates."

a. Tokenizing, Stoplist/Stopword/ Filtering, Stemming

Nama : "melon rock"

Abstrak : "melon rock melon rock rupa buah lokal buah golong buah sejati tunggal daging kulit luar buah warna hijau ke abu daging buah warna oranye hermanto 2013 empat camat jadi sampel melon rock hanya tua camat kaliwates".

3.2.2 Perhitungan TF-IDF

Tabel 3.1 Pembobotan TF*IDF

TERM	Q	TERM FREQUENCY				IDF	TF-IDF			
		D1	D2	D3	DF		Q	D1	D2	D3
2013	0	1	1	1	3	1	0	0,4239	0,4239	0,4239
abu	0	0	0	1	1	1,47712	0	0	0	0,62615
ambulu	0	0	1	0	1	1,47712	0	0	0,62615	0
balung	0	0	1	0	1	1,47712	0	0	0,62615	0
buah	**1**	5	5	4	3	1	0,4239	2,11895	2,11895	1,69056
camat	0	2	1	1	3	1	0	0,84878	0,4239	0,4239
daging	**1**	2	2	1	3	1	0,4239	0,84878	0,84878	0,4239
empat	0	1	1	1	3	1	0	0,4239	0,4239	0,4239
gadung	0	4	0	0	1	1,47712	0	2,50461	0	0
golong	0	1	1	1	3	1	0	0,4239	0,4239	0,4239
hermanto	0	1	1	1	3	1	0	0,4239	0,4239	0,4239
hijau	0	1	1	1	3	1	0	0,4239	0,4239	0,4239
kaliwates	0	1	1	1	3	1	0	0,4239	0,4239	0,4239
kulit	0	1	1	1	3	1	0	0,4239	0,4239	0,4239
lokal	0	1	1	1	3	1	0	0,4239	0,4239	0,4239
luar	0	1	1	1	3	1	0	0,4239	0,4239	0,4239
manalagi	**1**	0	4	0	1	1,47712	0,6262	0	2,50461	0
mangga	**1**	4	4	0	2	1,17609	0,4985	1,99418	1,99418	0
melon	0	0	0	3	1	1,47712	0	0	0	1,87846
oranye	0	1	1	0	2	1,17609	0	0,49855	0,49855	0
rock	0	0	0	4	1	1,47712	0	0	0	2,50461
rupa	0	1	1	1	3	1	0	0,4239	0,4239	0,4239
sampel	0	1	1	1	3	1	0	0,4239	0,4239	0,4239
sejati	0	1	1	1	3	1	0	0,4239	0,4239	0,4239
tanggul	0	1	0	0	1	1,47712	0	0,62615	0	0
tua	0	1	1	0	2	1,17609	0	0,49855	0,49855	0
tunggal	0	0	0	1	1	1,47712	0	0	0	0,62615
warna	0	1	1	1	3	1	0	0,4239	0,4239	0,4239

Keterangan : Q (Query), D1 (Documen 1), DF (Document Frequercy), IDF (Invers Documen Frequency).

3.2.3 Perhitungan *Vecktor Space Model*

Berikut langkah *retrieval* menggunakan metode *Vector Space Model* menggunakan data pada tabel perhitungan TF-IDF :

a. Mengkalikan bobot antara bobot *query* dengan bobot *term* pada setiap dokumen

Q.D1 = (0*0,4239)+(0*0)+……. = 2,252037246

Q.D2 = (0*0,4239)+(0*0) +……. = 3,820307994

Q.D3 = (0*0,4239)+(0*0,62615)+……. = 0,898460028

b. Menghitung panjang *query* dengan mengakarkan jumlah kuadrat bobot *query.*

$|Q| = \sqrt{0^2 + 0^2 + \ldots} = 1$

c. Menghitung panjang dokumen dengan mengakarkan jumlah dari bobot dokumen yang dikuadratkan pada setiap dokumen.

$|D1| = \sqrt{(0^2 + 0^2 + \cdots) * (0{,}4239^2 + 0^2 + \ldots)} =$ 4,405094

$|D2| = \sqrt{(0^2 + 0^2 + \cdots) * (0{,}4239^2 + 0^2 + \ldots)} =$ 4,388376

$|D3| = \sqrt{(0^2 + 0^2 + \cdots) * (0{,}4239^2 + 0{,}62615^2 + \ldots)} =$ 4,019487

Menghitung *cosine similarity* dengan membagi bobot $Q.D_n$ dengan hasil perkalian antara panjang *query* ($|Q|$) dan panjang dokumen ($|D_n|$).

$\text{Cos}(Q, D1) = \frac{2{,}252037246}{(1 * 4{,}405094)} = 0{,}51123479$

$\text{Cos}(Q, D2) = \frac{3{,}820307994}{(1 * 4{,}388376)} = 0{,}87055165$

$\text{Cos}(Q, D3) = \frac{0{,}898460028}{(1 * 4{,}019487)} = 0{,}22352602$

Memiliki ambang batas (*treshold*) 0,53510415 yang di dapat dari jumlah total cosin di masing-masing dokumen dan di bagi banyak dokumen yang memiliki nilai kosinus lebih dari 0,0. sehinggah tingkat similaritas dari scenario pada tabel di atas yaitu D2, D1, D3.

DAFTAR PUSTAKA

1. Sawitri, Komarayanti. 2017. *Ensiklopedia Buahb-buahan Lokal Jember Berbasis Potensi Alam Jember.* p-ISSN 2527-7111; e-ISSN 2528-1615. http://jurnal.unmuhjember.ac.id/index.php/BIOMA/article/download/591/470 Di akses pada tanggal 15/04/2018.
2. Ristu, Saptono. 2014. *Pemanfaatan Vektor Space Model dan Metode Cosine Similarity Pada Fitur Deteksi Hama dan Penyakit Tanaman Padi.* https://www.researchgate.net/profile/Ristu_Saptono/publication/309544361_Pemanfaatan_Metode_Vector_Space_Model_dan_Metode_Cosine_Similarity_pada_Fitur_Deteksi_Hama_dan_Penyakit_Tanaman_Padi/links/58b65463aca27261e5166537/Pemanfaatan-Metode-Vector-Space-Model-dan-Metode-Cosine-Similarity-pada-Fitur-Deteksi-Hama-dan-Penyakit-Tanaman-Padi.pdf Di akase pada tanggal 15/04/2018.
3. Andi Librian. 2016. *Stemming Bahasa Indoensia.* https://github.com/sastrawi/sastrawi/wiki/Stemming-Bahasa-Indonesia Di akses pada tanggal 17/04/2018.
4. Dataq, 2013. *Perbedaan Recal, Precision dan Accuracy.* https://dataq.wordpress.com/2013/06/16/perbedaan-precision-recall-accuracy/ Di akses pada tangal 18/04/2018.