PEMERINGKATAN HASIL PENCARIAN DOKUMEN TEKS PADA MESIN PENCARI

by 08032023 Purwatiningtyas

Submission date: 08-Mar-2023 02:46PM (UTC+0700)
Submission ID: 2031964516
File name: 6._Jurnal_PEMERINGKATAN_HASIL_PENCARIAN_DOKUMEN_Sinta_5.pdf (520.13K)
Word count: 3183
Character count: 20018

PEMERINGKATAN HASIL PENCARIAN DOKUMEN TEKS PADA MESIN PENCARI

Fatkhul Amin[1], Purwatiningtyas[2], Jeffri Alfa[3]
[1,3]Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Stikubank
[2]Program Studi Sistem Informasi, Fakultas Teknologi Informasi, Universitas Stikubank
e-mail: [1]fatkhulamin@edu.unisbank.ac.id, [2]purwati@edu.unisbank.ac.id, [3]mrjf@edu.unisbank.ac.id

ABSTRAK

Mesin pencari dokumen teks berbahasa Jawa krama sukar ditemukan baik dari dunia daring maupun luring. Pencarian dokumen teks bahasa jawa krama dengan menggunakan mesin pencari yang tersedia didapatkan suatu hasil pencarian dengan dokumen terambil yang banyak (recall tinggi) sehingga hasil menjadi kurang akurat (precision rendah). pemeringkatan mesin pencari dengan metode DICE Similarity agar user dibuat untuk mempermudah dalam melakukan pencarian dokumen teks berbahasa Jawa krama. Metode DICE Similarity akan bisa menghitung dan menampilkan dokumen yang ada pada database sehingga antar dokumen memiliki bobot yang berbeda untuk menentukan dokumen mana yang paling mirip (similar) dengan query. Dokumen teks yang terletak di posisi atas akan ditempati oleh dokumen dengan bobot tertinggi hasil pencarian metode DICE. Evaluasi hasil pencarian Mesin pencari dilakukan dengan uji recall dan precision dengan model persepsi. Mesin pencari mampu melakukan pencarian dokumen dan menampilkan hasil pencarian dokumen memiliki rata-rata recall 0,03 dan rata-rata precision 0,79.

Kata-Kunci: *Mesin Pencari*, Basa Jawa krama, *Dice Similarity*

1. PENDAHULUAN

Bangsa Indonesia adalah bangsa yang berbudaya dan memiliki ratusan etnis yang didalamnya juga memiliki ratusan bahasa daerah. Bahasa Jawa sebagai bahasa yang paling banyak digunakan di wilayah Indonesia setelah bahasa Indonesia, dewasa ini mulai banyak ditinggalkan oleh kebanyakan orang, terutama Bahasa Jawa Krama (halus). Media offline dan media online juga kurang mengangkat bahasa Jawa Krama sehingga dikhawatirkan bahasa Jawa Krama lama-kelamaan akan ditinggalkan oleh bangsa kita. Beberapa media online berbahasa Jawa ada, namun sedikit yang membahasa bahasa jawa krama dan belum menggunakan atau belum menyediakan pencarian informasi menggunakan mesin pencari khusus berbahasa jawa. Bahasa daerah adalah bahasa yang terkait akan latar belakang etnis, suku, budaya, yang begitu kaya di Indonesia. Bahasa daerah mencerminkan identitas bangsa ini, cermin kita sebagai bangsa yang kaya akan budaya dan bahasa. Bangsa Indonesia memiliki sekitar 700 lebih bahasa daerah, tetapi yang tercatat oleh Kementrian Pendidikan dan Kebudayaan (Kemendikbud) hanya sekitar 470 saja.

Ada 2 (dua) faktor utama yang menyebabkan bahasa Jawa krama (bahasa daerah pada umumnya) ditinggalkan oleh masyarakat, yaitu faktor internal dan faktor eksternal. Adapun Faktor internal yang dimaksud; 1) Melemahnya Sosialisasi dalam Keluarga, 2) Disorientasi Kurikulum Pendidikan, dan 3) Kurangnya Kesadaran Generasi Muda. Sedangkan Faktor eksternal yang menjadi penyebabnya yaitu; 1) Modernisasi dan Globalisasi, 2) Eksistensi Bahasa Asing di Indonesia, dan 3) Dominasi Kultural.

Pencarian informasi saat ini dilakukan dengan menggunakan Mesin Pencari atau mesin pencari, *user* menuliskan *query* dan mesin pencari akan menampilkanlkan hasil pencarian. 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 pengguna (*user)*. Hal ini terjadi karena dokumen yang terambil oleh sistem jumlahnya banyak, maka sistem berkemungkinan menampilkanlkan hasil pencarian yang tidak relevan. Banyaknya dokumen hasil pencarian ini membuat waktu yang dibutuhkan dalam pencarian menjadi lebih banyak dari yang diharapkan.

2. TINJAUAN PUSTAKA

Penelitian terkait dengan menggunakan metode *DICE Similarity* dilakukan antara lain pada bidang *Computer Science,* (Manoj Chahal, 2016) Ada Volume data yang besar yang tersedia di dunia digital. Untuk mengambil informasi yang relevan dan berguna dalam dunia digital ini merupakan salah satu tugas yang menantang. Tidak mungkin untuk pengguna untuk mengambil data yang efisien secara manual. Untuk mengatasi ini mesin pencari masalah digunakan. Cari menggunakan mesin Sistem Retrieval Informasi dan Algoritma Genetika untuk mengambil informasi yang relevan. Dalam makalah ini Dice Kesamaan Fungsi digunakan untuk mengambil informasi yang relevan .Dice kesamaan digunakan dalam Algoritma genetik untuk mendapatkan data yang efisien dari dunia digital.

Penelitian tentang *Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web* dilakukan oleh Vikas Thada, Dkk, (2015). Sebuah koefisien kemiripan mewakili kesamaan antara dua dokumen, dua pertanyaan, atau satu dokumen dan satu query. Dokumen yang diambil juga dapat peringkat di urutan dianggap penting. Sebuah koefisien kemiripan adalah fungsi yang menghitung tingkat kesamaan antara sepasang objek teks. Ada sejumlah besar koefisien kesamaan diusulkan dalam literatur, karena yang terbaik ukuran kesamaan tidak ada. Dalam makalah ini kami melakukan analisis perbandingan untuk mencari tahu dokumen yang paling relevan untuk himpunan kata kunci dengan menggunakan tiga koefisien kesamaan yaitu Jaccard, Dice dan Cosine koefisien. kita ini perform menggunakan pendekatan algoritma genetika. Karena sifat acak dari algoritma genetika yang terbaik nilai fitness adalah rata-rata 10 berjalan dari kode yang sama untuk tetap jumlah koefisien kemiripan iterations.The untuk satu set dokumen diambil untuk query yang diberikan dari Google yang mengetahui kemudian rata-rata relevansi dalam hal nilai-nilai fitness menggunakan koefisien kesamaan dihitung. Dalam makalah ini kami memiliki rata-rata 10 generasi yang berbeda untuk setiap query dengan menjalankan program 10 kali untuk nilai tetap Probabilitas Crossover Pc = 0,7 dan Probabilitas Mutasi Pm = 0,01. Percobaan yang sama dilakukan untuk 10

Penelitian tentang DICE Similarity juga dilakukan oleh Khuat Thanh Tung Dkk, (2015) mengukur kesamaan dokumenmemainkan peran penting dalam teks terkait penelitian danaplikasi seperti dokumen clustering, deteksi plagiarisme,pencarian informasi, terjemahan mesin dan esai otomatisscoring. Banyak penelitian telah diusulkan untuk memecahkan inimasalah. Mereka dapat dikelompokkan menjadi tiga pendekatan utama:String berbasis, Corpus-based dan Pengetahuan berbasis Persamaan.Dalam tulisan ini, kesamaan dua dokumen yang diukur denganmenggunakan dua ukuran berbasis-string yang berbasis karakter danalgoritma berbasis jangka. Dalam metode berbasis karakter, n-gram adalahdimanfaatkan untuk mencari sidik jari untuk sidik jari dan menampilkanalgoritma, maka koefisien Dice digunakan untuk mencocokkan duasidik jari yang ditemukan. Dalam pengukuran berbasis jangka, cosinusalgoritma kesamaan digunakan. Dalam karya ini, kami inginmembandingkan efektivitas algoritma yang digunakan untuk mengukurkesamaan antara dua dokumen. Dari hasil yang diperoleh,kita dapat menemukan bahwa kinerj sidik jari dan menampilkanlebih baik dari kesamaan kosinus. Selain itu, menampilkan yangalgoritma lebih stabil daripada yang lain.

Dice Similarity adalah metode untuk melihat tingkat kedekatan atau kesamaan *(smilarity)* term dengan cara pembobotan *term*. Dokumen dipandang sebagi sebuah vektor yang memiliki *magnitude* (jarak) dan *direction* (arah). Bobotistilahyangakhirnya digunakanuntuk menghitungtingkat kesamaanantarasetiap dokumenyang tersimpandalam sistem danpermintaan 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*). Dice similarity merupakan metode yangdigunakan untuk menghitug tingkat kesamaan(similarity) antar dua buah objek. Untuknotasi himpunan dapat digunakan rumus (1):

$$S_{Dice} = \frac{2\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2} \quad (1)$$

dimana p dan q adalah dokumen yang berbeda. p_i adalah term i yang ada di dokurnen p q_i adalah term i yang ada di dokumen q.

3. METODE PENELITIAN

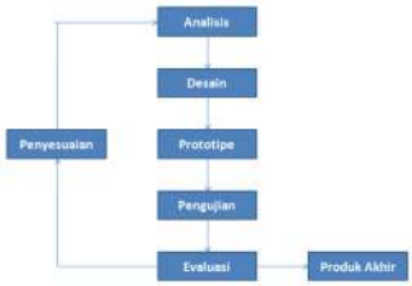
3.1 Pengumpulan data

Pengumpulan data dimaksudkan agar mendapatkan bahan-bahan yang relevan, akurat dan *reliable*. Maka teknik pengumpulan data yang dilakukan dalam penelitian ini adalah sebagai berikut :

a. Observasi
   Melakukan pengamatan dan pencatatan secara sistematis tentang hal-hal yang berhubungan dengan basis data dokumen teks bahasa jawa yang sesuai kebutuhan. Melakukan pengamatan materi bahasa Bahasa Jawa krama dan menganalisis Struktur Bahasa Jawa krama pada majalah Penjebar Semangad.
b. Studi Pustaka
   Pengumpulan data dari bahan-bahan referensi, arsip, dan dokumen yang berhubungan dengan permasalan dalam penelitian ini. Studi pustaka dilakukan dengan cara online dengan melihat web dan video-video Bahasa Jawa.

3.2 Metode Pengembangan

Penelitian ini menggunakan model *prototype*. Di dalam model ini sistem dirancang dan dibangun secara bertahap dan untuk setiap tahap pengembangan dilakukan percobaan-percobaan untuk melihat apakah sistem sudah bekerja sesuai dengan yang diinginkan. Sistematika model *prototype* terdapat pada Gambar 1 memperlihatkan tahapan pada *prototype*.



Gambar 1. prototipe

Berikut adalah tahapan yang dilakukan pada penelitian ini dengan metode pengembangan *prototype*

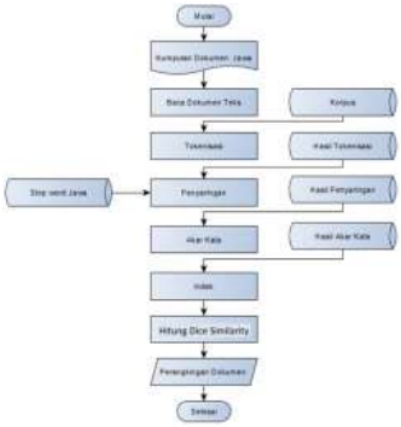
a. Analisa
   Pada tahap ini dilakukan analisa tentang masalah penelitian dan menentukan pemecahan masalah yang tepat untuk menyelesaikannya. Menentukan tujuan pembuatan mesin pencari.
b. Disain
   Pada tahap ini dibangun rancangan Sistem Temu Kembali Informasi bahasa jawa (DFD dan Flow Chart)
c. *Prototype*
   Pada tahap ini dibangun Sistem Temu Kembali Informasi Bahasa Jawa. Tahap ini di mulai dari proses tokenisasi, Penyaringan (filtering), Pembuatan kata dasar bahasa jawa (stemming), tfidf, dan perhitungan *Dice Similarity* yang diaplikasikan dengan program PHP.
d. Pengujian
   Pada tahap ini dilakukan pengujian *Recal* dan *Precission* dengan model Persepsi
e. Evaluasi
   Pada tahap ini dilakukan evaluasi apakah performa aplikasi sudah sesuai dengan yang diharapkan, apabila belum maka dilakukan penyesuaian-penyesuaian secukupnya.
f. Penyesuaian
   Tahap ini dilakukan apabila pada evaluasi performa aplikasi kurang memadai dan dibutuhkan perbaikan, tahap ini melakukan penyesuaian dan perbaikan pada aplikasi sesuia dengan kebutuhan

4. HASIL DAN PEMBAHASAN

4.1 Mesin Pencari Dokumen Teks Bahasa Jawa Krama

a. Flow Chart Mesin Pencari

Flowchart Mesin Pencari dokumen teks bahasa jawa krama diawali dengan *input* dokumen-dokumen kedalam korpus (gambar 2). Selanjutnya dokumen melalui proses preprosesing, dihitung bobotnya dan dibuat rankingnya berdasarkan bobot dokumen yang tertinggi.



Gambar 2. *Diagram Alir* STKI.

b. Tabel

Pada STKI ini menggunakan beberapa tabel untuk tempat meletakkan kumpulan data pada korpus, *term-term* hasil proses Tokenizing, *Filtering* dan *Stemming*. Selanjutnya untuk proses perhitungan dan pemeringkatan ***Dice Similarity*** digunakan tabel *freq* yaitu tabel yang berisi kumpulan *term-term* yang telah menjadi kata dasar. Berikut ini Rancangan tabel yang akan digunakan dalam STKI pada penelitian ini;

c. *Tampilan* Mesin Pencari Jawa Krama

Pada Aplikasi *interface* ini akan ditampilkan kolom *query* yang bisa digunakan untuk memasukkan *query* oleh pengguna. Kotak *button* dengan label cari digunakan untuk memproses setelah *query* di *input*. Tombol *button* cari jika sudah diklik akan menampilkan abstraksi hasil pencarian

4.2 PEMBAHASAN

a. Implementasi Mesin Pencari Bahasa *Jawa* metode *Dice Similarity*

Dokumen abstrak di *input* dengan cara manual dengan format dokumen teks. Proses ini dilakukan dengan cara memasukkan abstrak-abstrak skripsi bahan kajian penelitian kedalam tabel korpus. Sebelum dimasukkan kedalam tabel, dibuat satu tabel dengan nama tabel korpus yang digunakan sebagai tempat data. Tabel korpus ini memiliki *fiel-field* id, judul, isi dan dokumen. *Field* id berisi urutan data penelitian didalam korpus yang tersusun sesuai dengan urutan input data. Field judul berisi judul skripsi. Field isi berisi abstrak skripsi dan field dokumen berisi nama dokumen dengan kode tertentu. Proses memasukkan dokumen ke dalam tabel korpus ini memerlukan waktu relative lama bergantung pada jumlah data yang akan di *input* kedalam tabel korpus (Tabel 1).

Tabel 1. Tabel Korpus

b. Proses Tokenizing

Proses *scanner* dokumen korpus menggunakan format teks dilakukan dengan cara masuk kedalam dokumen korpus melalui perantara program php ke dalam database mysql. Proses *scanner* data dilakukan dengan cara *scanner* baris per baris, untuk tiap-tiap file naskah yang ada di dokumen. Tokenizing dimulai dengan memisahkan *term-term* yang ada pada dokumen korpus menjadi kumpulan term melalui proses *scanner* dengan dasar spasi. Selanjutnya term hasil proses Tokenizing di masukkan kedalam tabelawal dengan menyertakan *field-field* judul, *term* dan dokumen. Proses Tokenizing dilakukan dengan dua tahap yaitu tahap scanner *term* pada korpus kemudian term hasil *scanner* dimasukkan ke tabelawal dan tahap berikutnya adalah *scanner term* pada tabel awal dan menempatkan *term* hasil scanner di tabel kedua. Hasil *scanner file* pada proses (tabel 2)

Tabel 2. Tabel awal

judul	term	dokumen
Manfaat Daun Sirih	sinten	BK1
Manfaat Daun Sirih	ingkang	BK1
Manfaat Daun Sirih	mboten	BK1
Manfaat Daun Sirih	tepang	BK1
Manfaat Daun Sirih	kaliyan	BK1
Manfaat Daun Sirih	suruh?	BK1
Manfaat Daun Sirih	saking	BK1
Manfaat Daun Sirih	simbah-simbah	BK1
Manfaat Daun Sirih	kita	BK1
Manfaat Daun Sirih	sedaya	BK1
Manfaat Daun Sirih	tiyang	BK1
Manfaat Daun Sirih	indonesia	BK1
Manfaat Daun Sirih	kaliyan	BK1
Manfaat Daun Sirih	budaya	BK1
Manfaat Daun Sirih	"ninang	BK1
Manfaat Daun Sirih	suruh"	BK1

c. Proses *Filtering*

Proses selanjutnya setelah proses Tokenizing adalah proses *filtering*. Proses filtering dilakukan untuk menghilangkan *term-term* yang tidak memiliki arti dengan menggunakan *stopword list* tala. Proses filtering adalah proses baca tabel kedua untuk diperiksa apakah semua term memiliki term-term yang termasuk dalam stopword list menurut tala. Jika dalam tabel kedua terdapat *term-term* yang termasuk dalam *stopword*, maka akan dilakukan penghilangan *term-term* tersebut. Hasil proses filtering selanjutnya dimasukkan dalam tabel freq (tabel 3)

Tabel 3. Proses Filtering

judul	term	dokumen
Manfaat Daun Sirih	sinten	BK1
Manfaat Daun Sirih	ingkang	BK1
Manfaat Daun Sirih	mboten	BK1
Manfaat Daun Sirih	tepang	BK1
Manfaat Daun Sirih	kaliyan	BK1
Manfaat Daun Sirih	suruh?	BK1
Manfaat Daun Sirih	saking	BK1
Manfaat Daun Sirih	simbah-simbah	BK1
Manfaat Daun Sirih	kita	BK1
Manfaat Daun Sirih	sedaya	BK1
Manfaat Daun Sirih	tiyang	BK1
Manfaat Daun Sirih	indonesia	BK1
Manfaat Daun Sirih	kaliyan	BK1
Manfaat Daun Sirih	budaya	BK1
Manfaat Daun Sirih	"ninang	BK1
Manfaat Daun Sirih	suruh"	BK1

d. Proses *stopword Removal*

Proses membuang ***stopword (stopword removal)*** dilakukan untuk menghilangkan *term-term* yang tidak memiliki arti dengan menggunakan ***stopword jawa***. Proses ini dilakukan dengan cara *scanner* dan scanner tabel kedua

e. Proses *Stemming*

Proses *stemming* yang digunakan adalah proses *stemmer* menggunakan *stemmer* untuk bahasa Jawa ngoko berdasarkan stemmer bahasa Indonesia yang dibuat Tala. Proses stemming dengan menggunakan *stemmer* jawa melalui beberapa tahapan dan untuk mendukung proses ini juga digunakan *stopword list jawa*. Hasil akhir dari proses stemming adalah kumpulan *term* yang sudah menjadi kata dasar yang diinput dalam tabel *freq*. Proses *stemming* menghasilkan kumpulan term berupa kata dasar hasil scanner *term* pada tabel kedua. Proses *stemming* didukung stopword jawa yang digunakan untuk mengurangi term yang ada pada tabel kedua. Selanjutnya *term* hasil *stemming* di letakkan pada tabel freq

Tabel 4. Proses Stemming

judul	term	freq	freqpangkat
Manfaat Daun Sirih	sint	1	1
Manfaat Daun Sirih	ingkang	93	8649
Manfaat Daun Sirih	mbot	20	400
Manfaat Daun Sirih	tepang	4	16
Manfaat Daun Sirih	kaliy	35	1225
Manfaat Daun Sirih	suruh?	1	1
Manfaat Daun Sirih	saking	25	625
Manfaat Daun Sirih	simbah-simbah	1	1
Manfaat Daun Sirih	kita	16	256
Manfaat Daun Sirih	seday	18	324
Manfaat Daun Sirih	tiyang	19	361
Manfaat Daun Sirih	indonesi	11	121
Manfaat Daun Sirih	buday	5	25
Manfaat Daun Sirih	"ninang	1	1
Manfaat Daun Sirih	suruh"	1	1
Manfaat Daun Sirih	piyambak	7	49

f. Proses *Indexing*

Proses *indexing* dilakukan untuk mengambil atau meretrieve *term-term* yang ada pada tabelfreq untuk selanjutnya diproses pada saat pencarian dilakukan oleh STKI. Proses perhitungan dilakukan langsung pada STKI saat *query* diproses oleh sistem. User memasukkan Kata Kunci (*query*) pada mesin pencari, kemudian setelah kata kunci ditulis mesin pencari akan melakukan pencarian *query* pada *database* dengan mengolahnya terlebih dahulu sesuai dengan arsitektur mesin pencari menggunakan metode *vector space model* dan memberikan hasil pencarian.

g. Proses Perhitungan *Dice Similarity*

STKI metode Dice akan melakukan proses perhitungan dimulai dari menghitung tfidf, menghitung jarak *query* dan jarak dokumen, menghitung similaritas produk, dan menghitung bobot dokumen. STKI akan mengeksekusi *query* dari *user* dan akan mengolah *query* tersebut. *Query* yang di *input* oleh user selanjutnya akan dilakukan pencarian pada tabel freq kemudian dilakukan perhitungan pembobotan menggunakan metode ***Dice Similarity***. Perhitungan dilakukan dalam sistem pencarian, sistem pencarian akan melakukan perhitungan kemudian akan menampilkan hasilnya. Hasil pencarian akan menampilkan nama dokumen di korpus, kemudian bobot similaritas dan disusun berdasarkan perankingan. Bobot terbesar akan menempati ranking teratas pada hasi pencarian.

Dice Similarity adalah metode untuk melihat tingkat kedekatan atau kesamaan *(smilarity)* term dengan cara pembobotan *term*. Dokumen dipandang sebagi sebuah vektor yang memiliki *magnitude* (jarak) dan *direction* (arah). 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*).
Dice similarity merupakan metode yang digunakan untuk menghitug tingkat kesamaan (similarity) antar dua buah objek. Untuk notasi himpunan dapat digunakan rumus (1):

$$s_{Dice} = \frac{2\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2} \quad (1)$$

dimana p dan q adalah dokumen yang berbeda. p_i adalah term i yang ada di dokumen p q_i adalah term i yang ada di dokumen q.

h. Aplikasi STKI

Mesin Pencari Dice dirancang agar *user* mudah menggunakan dalam mencari dokumen yang relevan. Tampilan (*interface)* juga dirancang seperti mesin pencari pada umumnya, sehingga siapapun usernya akan langsung mudah beradaptasi dalam menggunakan mesin pencari. Prosedur menggunakan STKI ini sangat mudah, yaitu *user* hanya perlu menuliskan *query* atau kata kunci yang akan di cari pada kotak dialog kemudian setelah *query* di masukkan *user* tinggal mengklik tombol cari atau tekan *enter*.

Studi kasus pada aplikasi Mesin Pencari Dice ini menggunakan dokumen-dokumen Basa Jawa pada Majalah Online Penjebar Semangad yang terdapat pada 3 kategori yaitu; Kejawen, kebatinan, dan pasujarahan. *Query* yang dimasukkan pada Information Retrieval System adalah Studi kasus pada aplikasi mesin pencari ini menggunakan dokumen-dokumen teks berbahasa jawa krama. *Query* yang dimasukkan pada mesin pencari adalah *keyword* dengan 2 *term* yaitu "*wayang jawi*" , 3 *term* "*sejarah budaya wayang*", "budaya wayang jawi" 4 *term* "*sejarah budaya wayang wong*". 5 *term* "Cerita *sejarah budaya wayang wong*".

Gambar 3 Aplikasi STKI Jawa

Gambar 4 Hasil Pencarian *keyword*

Hasil pencarian dokumen dengan keyword "*wayang jawi*", menunjukkan dokumen dengan bobot tertinggi adalah dokumen letak dokumen BK9 (bobot 0,001).

i. Pengujian *recall* dan *precision*

Pengujian *recall (P)* dan *precision (R)* dilakukan dengan cara *input query* ke dalam Information Retrieval System *input* 1 *term*, 2 *term* dan 3 *term, 4 term, dan 5 term*. Perhitungan *recall* dan *precision* menggunakan persamaan (2) dan persamaan (3). Hasil pengujian *recall* dan *precision* dengan menguji 1 *term*, 2 *term* dan 3 *term* sampai dengan 5 term menunjukkan bahwa jika *recall* rendah maka *precision* akan tinggi, selengkapnya terlihat pada tabel 1.

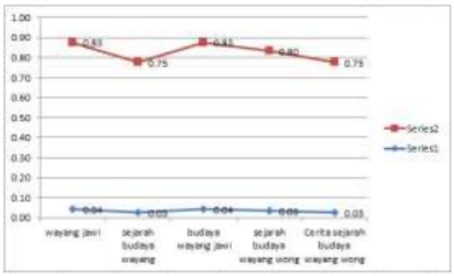
$$R = \frac{Number\ of\ relevant\ items\ retrieved}{Total\ number\ of\ relevant\ items\ in\ collection} \quad (2)$$

$$P = \frac{Number\ of\ relevant\ items\ retrieved}{Total\ number\ of\ items\ retrieved} \quad (3)$$

Tabel 5 Hasil Pengujian *Recall* dan *Precision*

No	Query	Recall	Precision
1	*wayang jawi*	0.04	0.83
2	sejarah budaya wayang	0.03	0.75
3	budaya wayang jawi	0.04	0.83
4	sejarah budaya wayang wong	0.03	0.80
5	Cerita sejarah budaya wayang wong	0.03	0.75

Hasil uji recall dan precision berdasarkan persepsi bisa dilihat pada gambar 5



Gambar 5 Diagram hasil uji Recall dan precision

5. SIMPULAN DAN SARAN

5.1. Kesimpulan

a. Mesin Pencari (information retrieval system) Dokumen Teks Bahasa Jawa Krama mampu melakukan pencarian dokumen teks bahasa jawa krama dan menampilkan hasil pencarian dokumen teks berbahasa Jawa dengan disertai bobot tiap dokumen beserta letak dokumen dengan metode DICE Similarity.
b. Hasil Uji *recall* dan *precision* mesin pencari menunjukkan hasil pencarian dokumen teks memiliki rata-rata *recall* = 0,03 dan rata-rata *precision* = 0,79.

5.2. Saran

a. Stemmer Jawa masih perlu perbaikan untuk Proses *stemming* perlu diperbaiki karena hasil yang didapatkan masih belum bisa sepenuhnya membuat semua *term* kedalam bentuk *term* kata dasar dengan benar.

DAFTAR PUSTAKA

[1] Budi, I., Aji, R.F., 2006. Efektifitas Seleksi Fitur dalam Sistem Temu Kembali Informasi. Seminar Nasional Aplikasi Teknologi Informasi (SNATI), ISSN : 1907-5022.

[2] Bum, K.Y., 2010. *An autonomous assessment system based on combined latent semantic kernels*. *Expert Systems with Applications: An International Journal , Volume 37 Issue 4*.

[3] Kadir, A., 2001. Dasar Pemrograman Web Dinamis menggunakan PHP. Penerbit Andi. Yogyakarta.

[4] Khuat Thanh Tung , (2016) A Comparison of Algorithms used to measure the Similarity between two documents, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 4, April 2015

[5] Manning, C., Raghavan, P., 2007. An Introduction to Information Retrieval, Stanford. USA.

[6] Manoj Chahal, 2016. Information Retrieval using Dice Similarity Coefficient , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 6, June 2016, ISSN: 2277 128

[7] Meadow, C.T., 1997. *Text Information Retrieval Systems*. Academic Press.New York.

[8] Tala, F.Z., 2003, *A Study of Stemming Effects on Information Retrieval in bahasa Indonesia*. Institut for logic, Language and Computation Universiteit van Amsterdam The Netherlands.

[9] R. Umamaheswari, K. Rajesh, 2014, Text Clustering Using Cosine Similarity and Matrix Factorization Cosine Similarity, International Journal of Research in Computer and Communication Technology, Vol 3, Issue 10, October – 2014

[10] Salton, G., 1989, *Automatic Text Processing, The Transformation, Analysis, and Retrieval of information by computer*. Addison – Wesly Publishing Company, Inc. USA.

[11] Vikas Thada , 2015. Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web, Department of Computer Science and Engineering Dr. K.N.M University,Newai, Rajasthan, India

[12] Yates, R.B, 1999. *Modern Information Retrieval, Addison Wesley*-Pearson international edition, Boston.USA

PEMERINGKATAN HASIL PENCARIAN DOKUMEN TEKS PADA MESIN PENCARI

ORIGINALITY REPORT

16%	15%	4%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

3%

★ eprints.umm.ac.id

Internet Source

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		