OPTIMASI PENCARIAN DATA MENGGUNAKAN TEXT FILTERING DAN ALGORITMA JARO WINKLER

Ardi Sanjaya

Program Studi Teknik Informatika, Fakultas Teknik, Universitas Nusantara PGRI Kediri
Kampus 2 Jl KH Ahmad Dahlan Gg1 No 6 Mojoroto Kediri jawa Timur
Email : dersky@gmail.com

Abstrak

Penelitian ini mencoba memberikan alternatif untuk optimasi pencarian data. Pembanding pengujian penelitian ini yaitu optimasi query yang menggunakan algoritma Jaro Winkler untuk pengurutan berdasar tingginya kemiripan. Data pengujian berupa deskripsi kondisi suatu produk diambil dari internet. Proses dimulai dengan memasukkan kata kunci pencarian. Kemudian dilakukan query dari database. Hasil dari query, kata yang tidak sama terhadap kata kunci pencarian dibuang (text filtering), lalu dilakukan pengukuran kemiripan menggunakan algoritma Jaro Winkler. Kemudian diurutkan berdasarkan nilai kemiripan terbesar. Kata kunci yang digunakan mulai dari 1 sampai 5 kata yang susunannya direkayasa mendekati data yang dikehendaki dan ada yang diacak. Secara keseluruhan berdasarkan hasil pengujian bahwa hasil optimasi pencarian data dengan membuang kata yang tidak sama saat pengukuran kemiripan memiliki peningkatan nilai kemiripan dan pada penyajian hasil query akan ditempatkan lebih atas (descending) berdasar nilai kemiripan yang besar. Kata kunci yang susunannya cenderung acak (susunan tertukar) dan tidak sama akan memiliki nilai kemiripan yang lebih rendah apabila tanpa melalui pembuangan kata yang tidak sama.

Kata kunci:*Optimasi Query, Pencarian Data, Algoritma Jaro Winkler.*

Abstract

This research tries to provide an alternative for data search optimization. Comparative testing of this research is query optimization which uses Jaro Winkler algorithm for sorting based on high similarity. Testing data in the form of a description of the condition of a product is taken from the internet. The process starts by entering search keywords. Then do a query from the database. The results of the query, words that are not the same as the search keywords are discarded (text filtering), then similarity measurements are performed using the Jaro Winkler algorithm. Then sorted according to the greatest similarity value. The keywords used range from 1 to 5 words whose arrangement is engineered to approach the desired data and some have been randomized. Overall based on the test results that the results of optimization of data search by removing words that are not the same when the similarity measurement has an increase in the similarity value and on the presentation of the query results will be placed higher (descending) based on a large similarity value. Keywords whose order tends to be random and not the same will have a lower similarity value if without going through the disposal of different words.

Keywords : *Query Optimization,Data Searching, Jaro Winkler Algorythm*

1. PENDAHULUAN

Pada masa sekarang, efektifitas dalam pencarian data masih menjadi topik yang terus diteliti dan dikembangkan. Bisa dibayangkan apabila pada suatu *database* yang besar namun sistem pencarian data masih kurang maksimal maka produktifitas akan menurun. Beberapa optimasi pencarian data yang telah dilakukan diantaranya dengan menguraikan kalimat yang dijadikan acuan pencarian menjadi kata kunci dan dikombinasikan dengan operator OR pada syarat pencarian. Dengan penguraian kalimat dan kombinasi operator OR pada syarat pencarian menjadikan cakupan pencarian lebih luas. Berdasarkan hasil, diperoleh bahwa pencarian data menggunakan penguraian kalimat menjadi kata kunci sebagai syarat pencarian lebih bisa menemukan data yang dimaksud [1]. [2] melakukan optimasi pencarian data dengan cara mengukur kemiripan kata kunci pencarian terhadap data hasil *query* dan mengurutkan data secara

descending berdasar nilai kemiripan menggunakan algoritma Jaro Winkler dan Levenstein Distance.

Penulis mengamati sistem pencarian barang pada beberapa sistem penjualan online berbasis web. Salah satu masalah yang dihadapi yaitu belum optimalnya sistem pencarian barang atau data ketika kata kunci pencarian lebih dari 1 kata dan susunan kata yang tertukar. Sistem pencarian pada beberapa sistem umumnya masih menggunakan sistem pencarian konvensional, yaitu pencarian yang mencocokkan kata kunci dengan data yang ada. Pencarian konvensional memiliki kelemahan yaitu tidak bisa mencari data yang relevan dengan kata kunci [3].

Penelitian ini bertujuan untuk mencari alternatif dalam optimasi pencarian data dengan menggunakan metode *text filtering* dan Jaro Winkler untuk mengukur kemiripan. Dimana dari nilai kemiripan yang tinggi akan diprioritaskan untuk tampil diawal (pengurutan secara *descending*). *Text filtering* yang dimaksud adalah pemisahan kata dan membuang kata yang tidak diperlukan.

2. DASAR TEORI

Data adalah sesuatu yang belum mempunyai arti bagipenerimanya dan masih memerlukan adanya suatupengolahan. Data bisa berwujud suatu keadaan, gambar, suara, huruf, angka, matematika, bahasa ataupun simbol-simbol lainnya yang bisa kita gunakan sebagai bahan untuk melihat lingkungan, obyek, kejadian ataupun suatu konsep [4].

Jaro-Winkler merupakan varian dari metrik Jaro Distance biasanya digunakan pada bidang keterkaitan rekaman (duplikat) dirancang dan paling sesuai untuk string pendek. Pada JaroWinkler untuk dua string semakin tinggi jarak, semakin mirip data yang diperoleh dengan skor 0 sama dengan tidak ada persamaan dan 1 sama persis [2].

Algoritma Jaro-Winkler *distance* memiliki kompleksitas waktu *quadratic runtime complexity* yang sangat efektif pada string pendek dan dapat bekerja lebih cepat dari algoritma *edit distance* [5]. Dasar dari algoritma ini memiliki tiga bagian:

a. Menghitung panjang string.
b. Menemukan jumlah karakter yang sama di dalam dua string.
c. Menemukan jumlah transposisi.

Pada algoritma Jaro-Winkler digunakan rumus untuk menghitung jarak (dj) antara dua string yaitu s1 dan s2 dapat dilihat pada persamaan (1) [4] :

$$d_j = \frac{1}{3} X \left(\frac{m}{s1} + \frac{m}{s2} + \frac{m-t}{s} \right) \quad (1)$$

Dimana m adalah jumlah karakter yang sama persis, |s1| adalah panjang string 1, |s2| adalah panjang String 2, T adalah jumlah transposisi, dan dj adalah Nilai jarak antara dua buah string yang dibandingkan.

Jaro-Winkler *distance* menggunakan *prefix scale* (p) yang memberikan tingkat penilaian yang lebih, dan *prefix length* (l) yang menyatakan panjang awalan yaitu panjang karakter yang sama dari *string* yang dibandingkan sampai ditemukannya ketidaksamaan. Bila *string* s1 dan s2 yang diperbandingkan, maka Jaro-Winkler *distance*-nya (dw) seperti pada persamaan (2) [5] :

$$dw = dj + (lp(1 - dj)) \quad (2)$$

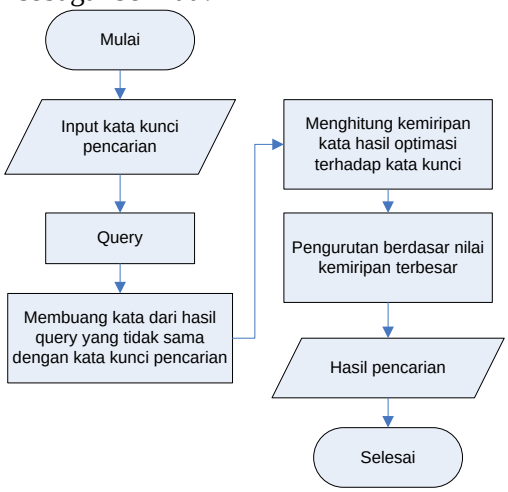
Dimana dw adalah nilai Jaro-Winkler *Distance*, dj adalah Jaro *distance* untuk strings s1 dan s2, l adalah panjang prefiks umum di awal *string* nilai maksimalnya 4 karakter (panjang karakter yg sama sebelum ditemukan ketidaksamaan max 4), dan p adalah konstanta *scaling factor* [5].

3. METODOLOGI PENELITIAN

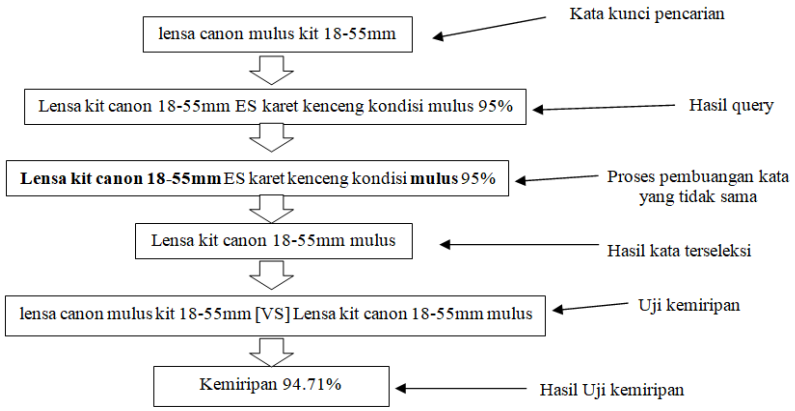
Pembanding pengujian penelitian ini yaitu optimasi *query* yang menggunakan algoritma Jaro Winkler untuk pengurutan berdasar tingginya kemiripan. Pada pengujian ini, peneliti bermaksud mencari suatu deskripsi dari produk dimana yang dikehendaki adalah produk dengan kode barang 1012 yaitu yang memiliki deskripsi “Lensa kit canon 18-55mm ES karet kenceng kondisi mulus 95%”. Data berupa deskripsi kondisi suatu produk diambil dari internet. Proses

dimulai dengan memasukkan kata kunci pencarian. Kemudian dilakukan *query* dari *database*. Hasil dari *query*, kata yang tidak sama terhadap kata kunci pencarian dibuang (*text filtering*), lalu dilakukan pengukuran kemiripan menggunakan algoritma Jaro Winkler. Kemudian diurutkan berdasarkan nilai kemiripan terbesar. Kata kunci yang digunakan mulai dari 1 sampai 5 kata yang susunannya direkayasa mendekati data yang dikehendaki dan ada yang diacak.

Alur optimasi pencarian data pada penelitian ini digambarkan pada gambar 1 dan diilustrasikan ada gambar 2 sebagai berikut :



Gambar 1. Alur optimasi pencarian data.



Gambar 2. Ilustrasi proses

4. HASIL DAN PEMBAHASAN

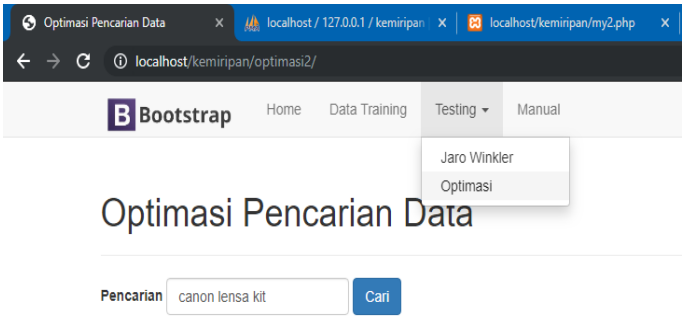
Data pengujian berisi judul dan deskripsi. Pada saat pemrosesan pencarian, data acuan yang digunakan adalah data pada kolom deskripsi. Data deskripsi produk yang digunakan pada penelitian ini disajikan pada tabel 1 sebagai berikut :

Tabel 1. Data Pengujian

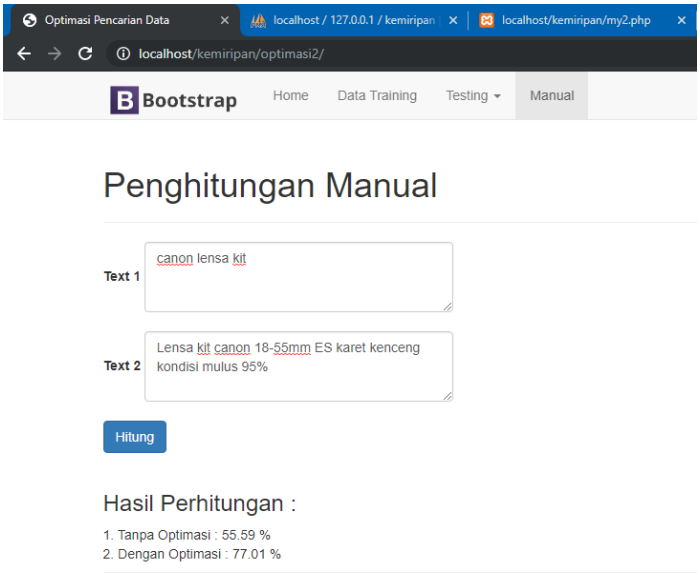
ID	JUDUL	DESKRIPSI
1001	OLYMPUS E-PL6 kit 14-42mm	Olympus E-PL6 Mirrorless Lensa Kit 14-42mm Box Optik bersih kondisi 95% mulus
1002	CANON EOS M10 kit 15-45mm	Canon EOS M10 Mirrorless Lensa kit 15-45mm Box lengkap Optik bersih kondisi 96% mulus
1003	LENSA OLYMPUS	LENSA OLYMPUS 14-42mm f3.5-5.6 EZ ED MSC kondisi 98% mulus

1004	LENSA CANON EF-S	LENSA CANON EF-S 18-135mm IS No Box Optik bersih Kondisi 97%
1005	Canon 600D Kit	Canon DSLR 600D lensa kit 18-55mm SC 9rb optik bersih kondisi 96%
1006	Canon 60D Body Only	Canon DSLR 60D Body Only BO SC 12rb karet masih kenceng kondisi 95%
1007	Lensa Sony 16-50mm	Lensa Sony 16-50mm PZ E Mount OSS optik bersih Rear Back Cap kondisi mulus 98%
1008	Sony A6000 BO	Sony A6000 mirrorless body only BO box kondisi mulus 98%
1009	Lensa wide Canon	Lensa Wide Mount canon 10-16mm Rear Back Cap kondisi mulus 97%
1010	Canon EOS 6D Mark II	Canon EOS 6D DSLR Mark II lensa Kit 24-105mm mulus kondisi 95%
1011	Lensa tamron tele nikon	Lensa tamron tele 70-300mm mount nikon kondisi 95% mulus
1012	Lensa kit 18-55mm canon	Lensa kit canon 18-55mm ES karet kenceng kondisi mulus 95%
1013	Lensa nikon kit 18-55mm	Lensa kit nikon 18-55mm VR karet kenceng kondisi mulus 98%
1014	Nikkor AF 80-200	Nikkor AF 80-200mm F2.8D ED Gen III rear back cap kondisi 90%
1015	Lensa Fix 85mm canon	Lensa canon Fix 85m F1.8 US Rear Back Cap Kondisi mulus 95%
1016	Kamera Nikon D90 Kit	DSLR Nikon D90 Kit 18-135mm Optik bersih kondisi 95%
1017	Lensa nikon 18-135mm	Lensa Nikon 18-135mm VR karet kenceng Rear Back Cap kondisi 95%
1018	Lensa Sony 18-200mm	Lensa sony 18-200mm E Mount F3.5-6.6 OSS Rear Back Cap kondisi 98%
1019	Lensa Canon 24-70mm	Lensa Canon 24-70mm EF F2.8 II USM Rear Back Cap kondisi 95%
1020	Lensa canon 24-105mm	Lensa canon 24-105mm EF F4L IS USM Rear Back Cap kondisi 96%

Tampilan aplikasi pada penelitian ini disajikan seperti pada gambar 3 dan 4 sebagai berikut :



Gambar 3. Tampilan Halaman Pengujian



Gambar 4. Penghitungan Manual

Peneliti melakukan 20 kali pengujian pencarian data dengan *keyword* yang berbeda. Didapati hasil seperti tersaji pada tabel 2 berikut :

Tabel 2. Hasil Pengujian (Dalam %)

No Pengujian	KeyWord	Jaro W	Optimasi	Selisih
1	lensa	87,82	100	12,18
2	lensa canon	89,2	100	10,8
3	lensa canon kit	86,11	96	9,89
4	lensa canon kit 18-55	88,64	92,19	3,55
5	lensa canon kit 18-55mm	89,35	97,39	8,04
6	lensa canon kit 18-55mm mulus	88,17	97,93	9,76
7	lensa canon 18-55	88,22	95,29	7,07
8	lensa canon 18-55mm	88,93	100	11,07
9	lensa canon 18-55mm kit	87,51	92,52	5,01
10	lensa canon mulus 18-55mm kit	85,44	92,11	6,67
11	lensa canon mulus 18-55mm	84,6	96,53	11,93
12	lensa canon mulus kit 18-55mm	85,51	94,71	9,2
13	canon lensa kit	55,59	77,01	21,42
14	canon lensa kit mulus	60,21	89,68	29,47
15	canon lensa kit mulus 18-55	62,17	82,28	20,11
16	canon lensa kit mulus 18-55mm	61,77	85,06	23,29
17	canon lensa mulus 18-55mm kit	60,82	78,38	17,56
18	mulus canon kit lensa	55,38	72,6	17,22
19	mulus canon lensa	46,07	70,21	24,14
20	mulus canon lensa kit	55,43	63,49	8,06

Pada tabel 2 diatas, pengujian nomor 1, kata kunci pencarian "lensa" memiliki nilai kemiripan 87,82% terhadap hasil *query* "Lensa kit canon 18-55mm ES karet kenceng kondisi mulus 95%" pada pengujian pembanding. Sedangkan untuk hasil optimasi dimana kata yang tidak sama dibuang pada hasil *query* menjadi "lensa" dan memiliki nilai kemiripan 100%. Dengan demikian hasil query akan ditampilkan paling atas (*descending*).

Pada pengujian nomor 5, kata kunci pencarian *"*lensa canon kit 18-55mm*"* memiliki nilai kemiripan 89,35% terhadap hasil *query* "Lensa kit canon 18-55mm ES karet kenceng kondisi

mulus 95%" pada pengujian pembanding. Sedangkan untuk hasil optimasi dimana kata yang tidak sama dibuang pada hasil query menjadi "lensa kit canon 18-55mm" dan memiliki nilai kemiripan 97,36%.

Pada pengujian nomor 13, kata kunci pencarian "canon lensa kit" memiliki nilai kemiripan 55,59% terhadap hasil query "Lensa kit canon 18-55mm ES karet kenceng kondisi mulus 95%" pada pengujian pembanding. Sedangkan untuk hasil optimasi dimana kata yang tidak sama dibuang pada hasil *query* menjadi "lensa kit canon" dan memiliki nilai kemiripan 77,01%.

Pada pengujian nomor 20, kata kunci pencarian "mulus canon lensa kit" memiliki nilai kemiripan 55,43% terhadap hasil *query* "Lensa kit canon 18-55mm ES karet kenceng kondisi mulus 95%" pada pengujian pembanding. Sedangkan untuk hasil optimasi dimana kata yang tidak sama dibuang pada hasil *query* menjadi "lensa kit canon mulus" dan memiliki nilai kemiripan 63,49%.

5. KESIMPULAN

Secara keseluruhan berdasarkan hasil pengujian bahwa hasil optimasi pencarian data dengan membuang kata yang tidak sama saat pengukuran kemiripan memiliki peningkatan nilai kemiripan dan pada penyajian hasil *query* akan ditempatkan lebih atas (*descending*) berdasar nilai kemiripan yang besar. Kata kunci yang susunannya cenderung acak (susunan tertukar) dan tidak sama akan memiliki nilai kemiripan yang lebih rendah apabila tanpa melalui pembuangan kata yang tidak sama.

Daftar Pustaka

[1] Sanjaya. A, "Optimasi query Untuk Pencarian Data Menggunakan Penguraian Kalimat", *in Proc. Semnasteknomedia 2016*, pp. 3.7-13, Februari 6, 2016

[2] Yulianingsih, "Implementasi Algoritma Jaro Winkler dan Levenstein Distance Dalam Pencarian Data Pada Database", in journal String, Vol 2 No 1, Agustus 2017

[3] Maskur, Rizky. A. F, "Implementasi Web Semantik Untuk Aplikasi Pencarian Tugas Akhir Menggunakan Ontologi dan Consine Similarity", in journal Nero, Vol 2 No 1, 2015

[4] http://kuliah.dinus.ac.id/edi-nur/sb1-7.html, diakses 23 September 2019

[5] Okta'mal. F, Saptono. R, Eko. S. M, "Jaro Winkler Distance dan Stemming Untuk Deteksi Dini Hama Dan Penyakit Padi", in proc Sesindo 2015, pp. 305-312, Nopember 2, 2015.