Analisis Dan Implementasi Mesin Focused Crawler Untuk Web Olahraga Dengan Algoritma Best First Search

Analysis and Implementation Focused Crawler Machine for Sport Web Using Best First Search Algorithm

Novita Debora.[1] (simanjuntaknovitadebora@gmail.com)
Eko Darwiyanto, ST.,MT[2] (warihmaharani@yahoo.com)
Erda Guslinar .P, ST[3] (erda.guslinar@gmail.com)

[1, 2, 3]**Fakultas Informatika – Institut Teknologi Telkom**
Jl. Telekomunikasi, Dayeuhkolot Bandung 40257 Indonesia

ABSTRAK

Saat ini halaman *web* bertambah sangat banyak dan berkembang begitu cepat dan menjadi salah satu sarana penyebaran informasi baik itu personal, sosial maupun komersial. Semakin banyak pula orang yang membutuhkan informasi mengenai topik-topik tertentu misalnya tentang olahraga namun mengalami kesulitan untuk mendapatkan informasi yang relevan. Untuk itu dibutuhkan *Web Crawler* khusus untuk membantu pengguna internet mencari halaman yang relevan. *Web crawler* sendiri adalah suatu program yang melakukan proses scanning ke semua halaman-halaman internet untuk dibuat indexnya dan mendukung sebuah *search engine*.

Berbeda dengan *crawler* yang dipakai oleh *search engine* komersial yang pada umumnya bertujuan untuk mengumpulkan halaman Web sebanyak mungkin, *focused crawler* (juga sering disebut dengan *topical crawler*) secara selektif menelusuri dan mengambil halaman Web yang relevan dengan topik tertentu. Dalam tugas akhir ini, digunakan *classifier Naïve Bayes* untuk membedakan halaman web olahraga dan bukan olahraga, serta menggunakan *Best First Search* sebagai algoritma penelusuran antrian. Pemilihan nilai yang terbaik dilakukan dengan membandingkan skor hasil perhitungan *Cosine Similarity*.

Ditunjukkan bahwa algoritma *best first search* dan *classifier Naïve Bayes* akan membantu menelusuri halaman yang yang relevan terlebih dahulu.

Kata kunci: focused crawler, web olahraga, naïve bayes, best first search

ABSTRACT

Currently, web pages are growing fast and evolving rapidly and become one of the means of dissemination of information by personal, social and commercial. The more people who need information on certain topics such as on sports, but find it difficult to obtain relevant information. That requires a Web Crawler specifically to help Internet users find relevant pages. Own Web crawler is a program that does the scanning process to all internet pages to be made indexnya and support a search engine.

Unlike the crawler that used by commercial search engines which generally aim to collect many web page, focused crawler (called as topical crawlers oft) browse and retrieve web pages relevant to a particular topic selectively. In this bachelor thesis, Naive Bayes classifier is used to distinguish the web page instead of sports and non sports web page, and using Best First Search as the crawling algorithm of the queue. Selection of the best value was done by comparing the calculation's results of Cosine Similarity.

Shown that the best-first search algorithm and the Naive Bayes classifier will help browse the relevant pages first.

Keywords: focused crawler, sports web, naïve bayes, best first search

1. PENDAHULUAN

Kebutuhan manusia akan informasi saat ini sangat tinggi, salah satunya adalah informasi akan dunia olahraga. Beberapa dari mereka menggunakan informasi olahraga sebagai kepentingan bisnis, hobi, tugas, dan kepentingan lainnya. Bagi sebagian orang yang mobilitas hidupnya tinggi, internet akan sangat dibutuhkan dan aplikasi web merupakan bagian yang tak terpisahkan dari hal tersebut. Namun mengingat banyaknya halaman web yang tersebar di dunia ini maka seringkali penggunaan *search engine* dengan *crawler* biasa tidak terlalu memberikan hasil yang cocok dengan kebutuhan informasi pengguna. Melihat kecenderungan ini, maka dibutuhkan suatu aplikasi *crawler* yaitu *Focused Crawler* atau sering disebut juga dengan *Topical Crawler* yang hanya men-*download* halaman web yang relevan dengan topik yang ingin dicari dan menghindari aktivitas *download* untuk halaman page lainnya yang tidak berkepentingan dengan topik tertentu. Dalam tugas akhir ini, halaman web yang akan diproses adalah halaman web olahraga.

Dalam Tugas Akhir ini digunakan *Naïve Bayes Classifier* untuk membedakan kelas-kelas halaman page yang relevan atau tidak dengan topik. Halaman web yang relevan dengan topik akan disimpan dalam koleksi antrian atau disebut juga dengan *frontier*. *Focused Crawler* akan memprediksi probabilitas kecocokan antara web yang ada di internet dengan topik yang dimaksud sebelum memulai untuk melakukan *downloading*. Jika halaman tersebut relevan, maka akan dilakukan ekstraksi ke *outgoing link*-nya untuk penelusuran ke level yang lebih dalam lagi. Proses ini akan terus berulang sampai antrian atau frontier telah habis (sudah mencapai batasan halaman maksimum yang ditetapkan sebelumnya). Proses penelusuran terhadap halaman web inilah yang disebut dengan *crawling stategy*. Banyak algoritma yang telah diimplementasikan dalam *crawling strategy* seperti *Breadth-First, Depth First Search, Best First Search, Best-N-First Search, PageRank, SharkSearch* maupun *InfoSpiders.* Pada Tugas Akhir ini, algoritma yang akan digunakan adalah *Best First Search* dimana hanya melakukan penelusuran terhadap node-node yang promising dengan rule-rule tertentu. Dalam makalah *Augmenting Focused Crawling using Search Engine Queries* disebutkan bahwa algoritma ini merupakan algoritma yang sedang banyak diteliti dan dikembangkan. Penentuan node-node yang promising dilakukan dengan menghitung skor dari masing-masing halaman web dengan menggunakan metode tertentu. Dalam tugas akhir ini, metode yang digunakan adalah *Cosine Similarity*.

Adapun performansi dari web crawler bergantung pada banyaknya link relevan yang terjaring, *Focused crawler* biasanya mengandalkan *search engine* yang umum digunakan sebagai penentu awal atau sering disebut dengan starting point. Skenario pengujian akan dilakukan dengan melakukan proses *crawling* ke internet langsung dan melakukan pembatasan halaman maksimum. Halaman web yang diproses berformat .html yang telah dihitung terlebih dahulu tingkat relevansinya dengan kategori olahraga. Halaman yang terklasifikasi dalam halaman web olahraga adalah halaman yang berisikan informasi mengenai olahraga apapun. Pada umumnya untuk mengevaluasi kualitas dari *Focused crawler* dapat dilhat dengan menghitung tingkat akurasi, *precision, recall dan F-Measure* yang berupa satuan persen. Akurasi menunjukkan tingkat keakuratan sistem melakukan pengelompokan, *precision* menunjukkan kemampuan sistem melakukan pengelompokan suatu kelas pada dokumen yang dikunjungi, *recall* menunjukkan kemampuan sistem melakukan pengelompokan suatu kelas pada kumpulan dokumen, sedangkan *F-measure* untuk mengukur kualitas dengan melibatkan *precision* dan *recall* itu sendiri.

2. DASAR TEORI

2.1 Internet Dan Web

Internet merupakan suatu sumberdaya elektronik yang paling besar di dunia, bisa disebut juga sebagai sebuah jaringan yang terdiri dari beratus-ribu komputer yang menjangkau seluruh bumi. Internet tercatat sebagai tempat publikasi elektronik terbesar. Kita dapat berbagi apa saja melalui internet dan semuanya bisa ter-*publish* di internet. Ada banyak metode yang bisa digunakan untuk mempublikasikan informasi di internet. Salah satunya adalah dengan World Wide Web (www). Web awalnya hanya digunakan para ilmuwan di Swiss untuk bertukar dan berbagi informasi yang pada akhirnya dirilis di internet sampai sekarang.

Pada dasarnya web dibangun dari sebuah hypertext atau hypermedia, dimana sebuah dokumen tunggal dapat ditambahkan sebuah *link* untuk menuju dokumen lainnya. Dokumen tersebut dapat berlokasi yang sama dengan dokumen awal atau bisa dimana saja di seluruh dunia dan pengguna dapat mengakses dokumen (bisa merupakan gabungan dari teks, gambar, suara, animasi dan tipe data lainnya) tersebut di waktu yang sama. Salah satu bahasa pemrograman yang digunakan untuk menulis dokumen web adalah Hyper Text Mark-up Languange (HTML) yang merupakan bagian dari Standard Generalized Markup Language (SGML). HTML memperkenalkan konsep "*structured text*" dimana setiap teks dari keseluruhan dokument ditandai dengan adanya *tags*. Ketika browser membuka sebuah halaman web, HTML akan mengurai masing-masing baris untuk melihat bagaimana harusnya teks tersebut ditampilkan.

2.2 Text Preprocessing

Text Preprocessing merupakan salah satu tahapan dalam text mining. Dalam text mining didefenisikan sebagai penggalian informasi yang bersumber dari sekumpulan dokumen atau teks. Preprocessing dilakukan untuk menentukan fitur-fitur yang nantinya akan mewakili dokumen atau teks tersebut seperti tokenizing, case folding, stemming,dan stopword.

2.3 Pembobotan

Beberapa metode pembobotan yang telah dikenal adalah TF (Term Frequency), IDF (Inverse Document Frequency) serta TF-IDF yang merupakan gabungan dari kedua pembobotan sebelumnya. Tugas akhir ini akan menggunakan metode pembobotan TF-IDF.

2.5.1 TF (Term Frequency)

Metode ini merupakan metode yang paling sederhana. TF menentukan bobot *term* pada suatu dokumen berdasarkan jumlah kemunculannya dalam dokumen tersebut. Nilai jumlah kemunculan suatu kata (*term frequency*) diperhitungkan dalam pemberian bobot terhadap suatu kata. Semakin besar jumlah kemunculan suatu *term* (tf tinggi) dalam dokumen, semakin besar pula bobotnya dalam dokumen atau akan memberikan nilai kesesuian yang semakin besar. TF yang digunakan adalah modifikasi dari TF normalisasi, yaitu:

$$TF(i,j) = \frac{f(i,j)}{maxterm\ (j)}$$

(2- 1)

Keterangan: *TF(i,j)* = nilai TF untuk setiap term
f(i,j) = frekuensi kemunculan term i dalam dokumen j
maxterm(j) = total kemunculan term dalam seluruh dokumen j.

2.5.2 IDF (Inverse Document Frequency)

Metode ini melakukan pengurangan dominansi *term* yang sering muncul di berbagai dokumen. Hal ini diperlukan karena *term* yang banyak muncul di berbagai dokumen, dapat dianggap sebagai *term* umum (*common term*) sehingga tidak penting nilainya. Sebaliknya faktor kejarangmunculan kata (*term scarcity*) dalam koleksi dokumen harus diperhatikan dalam pemberian bobot. Menurut Mandala (dalam Witten, 1999) 'Kata yang muncul pada sedikit dokumen harus dipandang sebagai kata yang lebih penting (*uncommon tems*) daripada kata yang muncul pada banyak dokumen. Pembobotan akan memperhitungkan faktor kebalikan frekuensi dokumen yang mengandung suatu kata (*inverse document frequency*).

$$IDF(j) = log\left(\frac{D}{df(i)}\right) + 1$$

(2- 2)

Keterangan: *IDF (j)* = nilai IDF untuk setiap term
D = banyaknya dokumen
df(i) = banyaknya term i yang muncul dalam dokumen

2.5.3 TF-IDF

Metode TF-IDF merupakan metode pembobotan *term* yang banyak digunakan sebagai metode pembanding terhadap metode pembobotan baru. Pada metode ini, perhitungan bobot *term* dalam sebuah dokumen dilakukan dengan mengalikan nilai TF *(Term Frequency)* dengan IDF *(Inverse Document Frequency)*.

$$TFIDF(i,j) = TF(i,j) * IDF\ (j)$$

(2- 3)

2.4 Information Retrieval

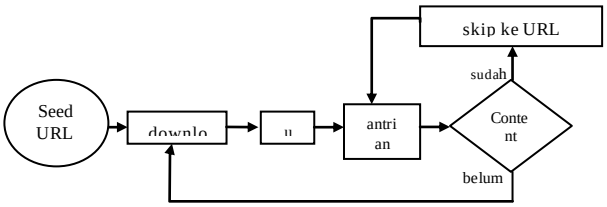
Information retrieval merupakan suatu seni dan ilmu untuk menemukan kembali informasi yang diinginkan user dari sekumpulan koleksi dokumen yang ada. Pada dasarnya *information retrieval* digunakan untuk sistem yang bersifat pribadi dimana hanya sebagian orang saja yang berkepentingan, misalnya mesin pencari dalam suatu perpustakaan besar. Namun seiring dengan berkembangnya zaman, orang-orang saling berhubungan dan saling membutuhkan informasi. Kondisi ini yang kemudian diadaptasi dalam sistem pencari yang lebih besar lagi contohnya web search engine (mesin pencari web). IR berhubungan dengan text, suara, gambar, ataupun data yang tidak selalu terstruktur dan ada kemungkinan kerancuan arti.

2.5 Web Crawler

Web crawler merupakan salah satu komponen dalam mesin pencari (seperti bing, yahoo, altavista, lycos, allfheweb dan atau search.msn, ataupun Google) yang bertugas untuk menjaring seluruh situs web yang ada (beberapa crawler melakukan secara periodik) yang selanjutnya akan dilakukan pengindeksan oleh mesin pencari tersebut. Masing-masing mesin pencari memiliki ciri tersendiri dalam pelacakan, misalnya: kecepatan pelacakan, ketepatan informasi, kuantitas situs pelacakan, teknik penelusuran, format dokumen yang dapat diakses dan lain sebagainya. Data tentang halaman web disimpan dalam sebuah database indeks untuk digunakan dalam pencarian selanjutnya (penelusuran periodik). Hal ini dikarenakan perkembangan halaman web yang sangat cepat. Sebagian mesin pencari, seperti Google, menyimpan seluruh atau sebagian halaman sumber

(yang disebut cache) maupun informasi tentang halaman web itu sendiri. Ketika seorang pengguna mengunjungi mesin pencari dan memasukkan query, biasanya dengan memasukkan kata kunci, mesin mencari indeks dan memberikan daftar halaman web yang paling sesuai dengan kriterianya, biasanya disertai ringkasan singkat mengenai judul dokumen dan terkadang sebagian teksnya. Manfaat mesin pencari bergantung pada relevansi hasil-hasil yang diberikannya. Meskipun mungkin ada jutaan halaman web yang mengandung suatu kata atau frase, sebagian halaman mungkin lebih relevan, populer atau autoritatif daripada yang lain.

Cara kerja crawler diawali dengan 1 atau lebih URL yang digunakan sebagai *data seed*. Dari URL ini crawler akan mengambil/ mendownload halaman web dan dimasukkan ke dalam koleksi kemudian diekstrak seluruh informasi dan link yang ada di dalamnya. Link (URL) yang telah diekstrak dan belum dikunjungi akan masuk dalam antrian. Hal ini akan terus berulang sampai seluruh URL dalam antrian telah habis atau telah memenuhi level yang ditetapkan. Untuk proses *crawling* yang berlanjut, URL yang telah ditelusuri akan kembali masuk dalam antrian untuk diurai kembali di periode berikutnya. Hal ini berguna untuk memastikan bahwa halaman tersebut masih aktif dan masih dikunjungi oleh pengguna.



Gambar 2-1 Web crawler secara umum

Terdapat 2 metode crawling yang utama:

1. *Exhautive crawling*

   Exhautive crawling bertujuan untuk mengumpulkan semua halaman web dengan melakukan strategi pencarian secara traversal, misalnya *Breadth First Search*. *Exhautive crawling* membutuhkan ruang penyimpanan yang sangat besar karena harus menelusuri seluruh web yang ada.
2. *Focused crawling*

   Berbeda dengan *exhautive*, *focused crawling* lebih bertujuan untuk mengumpulkan sebagian dokumen yang relevan dengan suatu topik tertentu dan mengabaikan *link* (URL) yang tidak berkaitan. *Focused crawler* biasanya diaplikasikan untuk suatu tujuan tertentu misalnya sebuah aplikasi kesehatan dimana hanya men-*download link-link* yang berhubungan dengan kesehatan.

Dalam menelusuri setiap URL (proses crawling), terdapat beberapa algoritma penelusuran. Beberapa diantaranya adalah *Breadth First Search*, *Depth First Search*, *Best First Search*, *Fish Search*, *Shark Search* dan lain-lain. Pada Tugas akhir ini algoritma yang digunakan adalah *Best First Search*.

2.5.1. Best First Search

Best First search merupakan algoritma penelusuran suatu masalah dengan mencari solusi terbaik. Halaman yang belum dikunjungi akan di-*ranking*sesuai skor yang diberikan dengan metode tertentu. Dalam tugas akhir ini metode yang digunakan adalah *cosine similarity*.

Berkaitan dengan *web crawler*, *best first search* biasanya digunakan dalam *focused crawler* dimana link-link diurutkan sesuai skor dahulu baru kemudian dimasukkan ke dalam antrian, sehingga yang menjadi *head* dari antrian adalah link dengan skor tertinggi. Dengan cara ini maka kita dapat memastikan bahwa halaman yang didownload sudah pasti memiliki tingkat kerelevanan yang paling tinggi.

2.5.2. Cosine Similarity

Cosine Similarity, dalam *information retrieval*, adalah metode yang telah banyak digunakan untuk menentukan nilai kemiripan antara suatu topik (query) dengan dokumen ataupun dokumen dengan dokumen. *Cosine similarity* akan menghitung dot product seperti di bawah ini:

$$\text{sim}\,(A,B) = \cos\theta\,(A,B) = \frac{|A \cap B|}{\sqrt{|A| \times |B|}}$$

(2- 4)

Keterangan:

A : document 1
B : document 2

Dimana:

$$\text{sim}\,(a,b) = \cos\theta\,(a,b) = \frac{\sum_k (a_{ik} . b_{jk})}{\sqrt{\sum_k a_{ik}^2}\sqrt{\sum_k b_{jk}^2}}$$

(2- 5)

Keterangan:

a_{ik} = term frequency untuk setiap term dalam dokumen 1

b_{jk}=term frequency untuk setiap term dalam dokumen 2

2.5.3 Classifier

Classification (klasifikasi) merupakan proses menentukan kelas (label) dari suatu objek yang tidak memiliki label. Pelabelan objek dilakukan berdasarkan kesamaan karakteristik antara sekumpulan objek (training set) dengan objek baru

tersebut. Teknik klasifikasi yang telah cukup dikenal saat ini adalah Naïve Bayes Classifier, Support Vector Machine, Case Based Reasoning, Genetic Algorithm dan masih banyak lagi. Dalam focused crawler, classifier berguna untuk mengklasifikasi halaman web sesuai dengan topik yang telah ditentukan. Cara kerja classifier dalam focused crawler terlihat ketika proses pelabelan terhadap kelas yang bukan olahraga dan kelas olahraga. Dalam Tugas akhir ini, classifier yang digunakan adalah Naïve Bayes classifier.

2.5.3.1 Teorema Bayes

Ide dasar dari teorema bayes adalah ketika kita dihadapkan pada masalah yang sifatnya hipotesis dimana mendesain fungsi klasifikasi untuk memisahkan 2 jenis objek, dalam tugas akhir ini adalah web olahraga dan web bukan olahraga.

$$P\ (A|C) = \frac{P(C|A)P(C)}{P(A)}$$

(2- 6)

Diketahui bahwa C merupakan data sample yang belum memiliki label kelas apapun, sedangkan A merupakan data yang telah memiliki label. Prior probability adalah probabilitas yang diperoleh dari sample data yang memilki label kelas A, sedangkan posterior probability adalah probabilitas label kelas A jika diketahui C.

P(A) = prior probability dari kelas A
P(C) = prior probability dari sample C
P(A|C) = posterior probability A dari sample C

2.5.3.2 Naïve Bayes Classifier

Merupakan salah satu metode classifier yang menerapkan teorema *Bayes*. Dengan memiliki asumsi bahwa nilai atribut-atribut akan memiliki hubungan saling bebas satu sama lain dalam suatu atribut kelas tertentu yang *given*, dalam tugas akhir ini merupakan kelas olahraga maupun kelas non olahraga.
Persamaan yang digunakan dalam klasifikasi Naïve Bayes adalah:

$$\mathrm{P\ (Ai|K)} = \mathrm{P(K)} \prod_{i=1}^{n} \mathrm{P\ (Ai\ |K)}$$

(2- 7)

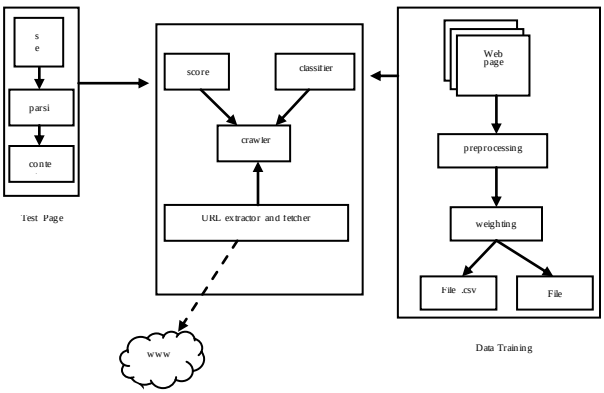
Dimana peluang posterior dihitung sesuai dengan teorema bayes sebelumnya, yaitu:

$$P\ (A|C) = \frac{P(C|A)P(C)}{P(C)}$$

(2- 8)

3. DESKRIPSI SISTEM

Sistem yang akan dibuat memiliki fungsi utama yaitu melakukan download URL yang telah terlebih dahulu ditentukan bobotnya untuk menjadi acuan penelusuran selanjutnya. Pertama-tama sistem akan melakukan penelusuran pada 3 halaman web uji (data seed) yang telah ditentukan terlebih dahulu dan bertopik olahraga. Hal ini dilakukan agar mesin crawler dapat melanjutkan penelusuran terhadap web bertopik olahraga. Setelah melakukan download content, crawler akan menyimpan halaman-halaman web tersebut ke antrian koleksi. Untuk halaman web yang mendapatkan urutan terdepan dalam antrian, dilakukan parsing pada halaman web sehingga ditemukan outgoing linknya berikutnya. Proses ini dinamakan ekstraksi informasi. Penelusuran dilakukan dengan algoritma Best First Search dimana penelusuran akan mengikuti urutan dimulai dari bobot yang tertinggi. Halaman-halaman yang ditelusuri adalah halaman yang terklasifikasi olahraga.Jika web tersebut merupakan web olahraga (dengan bobot tinggi), maka akan ditelusuri outgoing linknya dan akan masuk ke dalam antrian selanjutnya. *Classifier* yang digunakan adalah Naïve Bayes. Penelusuran akan dibatasi tingkat kedalamannya (level) dimana URL seed berada di level pertama. Proses download, parsing, klasifikasi, ranking akan dilakukan terus menerus sampai level maksimum yang telah ditentukan.



Gambar 3-1 Gambar arsitektur Focused Crawler

4. PENGUJIAN DAN ANALISIS

Data training yang digunakan berasal dari halaman web yang berbahasa Indonesia dengan kategori olahraga dan bukan olahraga. Data training yang digunakan tidak redundan sehingga pengetahuan ketika pembangunan model semakin banyak. Jumlah data training yang digunakan dalam proses crawling akan mempengaruhi hasil dari focused crawler itu sendiri. Tujuan dari pengujian

ini adalah untuk memperoleh hasil maksimal (evaluasi parameter pengujian) pada proses crawling. Jumlah data set yang digunakan sebanyak 300 dengan pembagian 50, 100, 150, 200, 250, 300.

4.1 Analisis terhadap akurasi sistem

Jika dilakukan pengujian pada halaman maksimum 50, maka akurasi yang didapatkan adalah sebagai berikut:

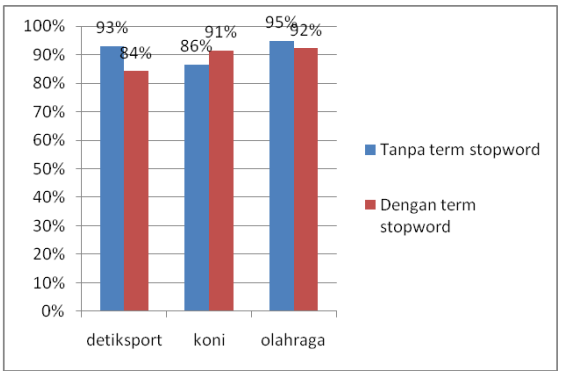
Data training	Akurasi rata-rata
50	87,22 %
100	91,39 %
150	76,97 %
200	77,08 %
250	74,32 %
300	78,01 %

Tabel 4-1 Akurasi dari link pengujian

Terlihat bahwa jika jumlah dataset yang digunakan dalam proses training ditambah (lebih dari 300) maupun ditambah (kurang dari 50), nilai akurasi dapat naik ataupun turun (bukan linear lurus). Perbedaan akurasi antar interval 50 pada dataset yang digunakan menghasilkan selisih akurasi $0{,}11\% \leq x \leq 14{,}42\%$.

Semakin banyak jumlah dataset dalam data training yang digunakan tidak menjamin bahwa akurasi akan semakin naik. Hal ini dikarenakan naïve bayes memiliki sifat independensi kondisional dimana kemunculan suatu term tertentu tidak mempengaruhi kemunculan term lainnya serta terdapat probabilitas bersyarat untuk setiap term sesuai dengan kelasnya. Dalam hal ini kelas olahraga dan non olahraga. Pemilihan term-term yang terlibat dalam data training juga sangat berpengaruh terhadap akurasi. Ketika suatu data training memiliki term-term penting lebih banyak (kata-kata yang dapat menjadi pembeda yang tepat), maka akurasi akan meningkat. Dalam hal ini, data training 100 memiliki lebih banyak koleksi term yang dapat dijadikan pembeda yang tepat, sehingga memberikan akurasi yang paling tinggi diantara data training lainnya. Hasil akurasi pada tabel diatas dapat naik ataupun turun jika dilakukan pada link pengujian yang berbeda. Hal ini dikarenakan perbedaan isi konten web. Walaupun secara aktual dikatakan sebagai web olahraga, namun halaman tersebut juga berisi konten-konten topik lain yang tidak berhubungan dengan olahraga.

4.2. Analisis terhadap pengaruh stopword

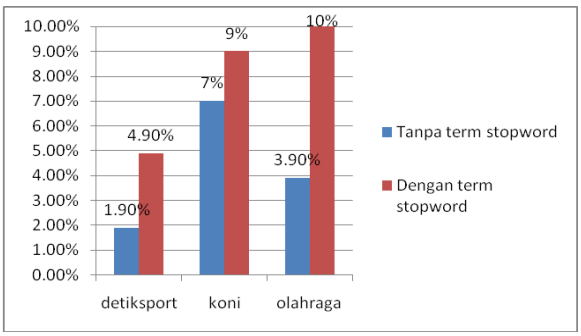


Gambar 4-1 Grafik akurasi untuk pengaruh stopword

Terlihat di grafik bahwa nilai akurasi hasil pengujian dataset tanpa menggunakan term stopword lebih tinggi dibandingkan dengan akurasi dengan menggunakan term stopword. Ini dikarenakan banyaknya term-term yang tidak menggambarkan kelas olahraga yang sebenarnya. Dalam hal ini, penghilangan kata-kata umum (stopword) sudah relatif baik karena nilai akurasi halaman pengujian yang dibandingkan dengan data training tanpa stopword terlihat lebih tinggi. Sehingga kata-kata yang terlibat dalam data training sudah dapat menjadi pembeda baik dalam pengujian.

Terdapat anomali pada pada link uji II dimana nilai akurasi pengujian dengan menggunakan term stopword lebih tinggi dibandingkan dengan tanpa menggunakan term stopword. Ini dikarenakan tidak banyak term-term yang mengandung term stopword pada pada halaman yang ditelusuri dalam link II sehingga akurasi nya lebih tinggi ketika diuji dengan data training dengan term stopword.

Skoring merupakan hal yang penting dalam penentuan link berikutnya yang akan ditelusuri. Penentuan skor dihitung berdasarkan cosine similarity yang telah dijelaskan pada sub bab II.



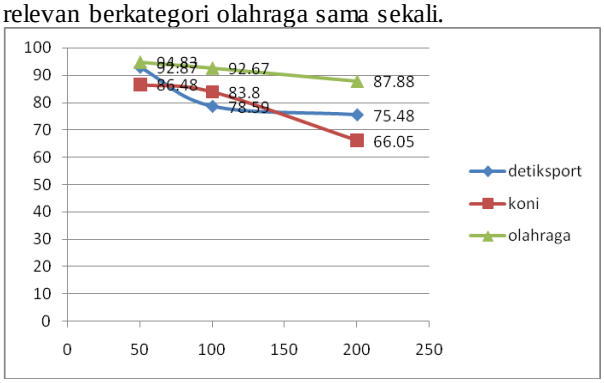
Gambar 4-2 Grafik skor kemiripan dokumen untuk pengaruh stopword

Terlihat bahwa term tanpa stopword memiliki skor yang lebih rendah, ini dikarenakan term-term yang berupa kata-kata umum tidak ikut dibandingkan dengan data testing sehingga menghasilkan skor yang lebih kecil. Skor yang dihasilkan dari perhitungan data training yang mengandung term stopword dengan data testing akan menghasilkan skor yang lebih tinggi karena memiliki banyak kesamaan antar setiap kata umum yang ada dalam data training maupun halaman testing. Persentase skor relatif rendah karena merupakan rata-rata dari seluruh dokumen yang telah dibandingkan antara data training dan halaman testing.

4.3 Analisis terhadap parameter pengujian

Precision yes, recall yes dan F-measure yes menandakan Precision, recall dan F-measure pada halaman web yang terkategori olahraga. Sedangkan Precision no, recall no dan F-measure no menandakan Precision, recall dan F-measure pada halaman web yang terkategori bukan olahraga.

Precision dan recall yes yang tinggi menunjukkan bahwa hasil focused crawler sudah relatif baik karena jumlah halaman-halaman yang relevan olahraga mendekati jumlah halaman-halaman yang terklasifikasi olahraga (precision) ataupun yang kelas aktualnya adalah olahraga (recall). Precision dan recall yes tertinggi berada pada halaman maksimum 50 dimana precision tertinggi mencapai 99,58% dan recall tertinggi mencapai 99,77%. Precision no dan recall no bernilai rendah karena hasil crawling yang ditelusuri jarang mendapatkan halaman yang berkategori tidak, bahkan ada yang tidak ada relevan berkategori olahraga sama sekali.



Gambar 4-3 Grafik akurasi hasil pengujian pada halaman maksimum

Link pengujian dengan maksimum halaman 50 memiliki akurasi yang relatif tinggi karena halaman penelusuran di awal (root) sudah merupakan kategori olahraga (tidak terklasifikasi salah), sehingga untuk 49 halaman berikutnya masih memiliki kemungkinan relevansi yang tinggi juga. Grafik akurasi terlihat menurun untuk setiap link pengujian jika maksimum halaman dinaikkan. Pada link uji II terlihat bahwa untuk 50 halaman maksimum, akurasi mencapai 86,48% dan menurun pada 100 halaman maksimum menjadi 83,8%. Penurunan akurasi ini juga terlihat pada 200 halaman maksimum menjadi 66,05%. Pola penurunan grafik akurasi ini juga terlihat pada link uji III. Ini dikarenakan crawler menjaring halaman-halaman yang lebih dalam yang memiliki tingkat relevansi yang semakin rendah dari halaman-halaman terdahulu dan memungkinkan classifier untuk melakukan kesalahan dalam klasifikasi halaman web. Ini akan berpengaruh ke link-link yang akan ditelusuri selanjutnya.

Pengujian yang dilakukan dengan link netral (bukan merupakan link olahraga) menghasilkan perbedaan akurasi yang tidak terlalu signifikan dengan seed yang merupakan web olahraga. Namun dari hasil link, terlihat banyak web bukan olahraga yang ikut terjaring. Hal ini terlihat pada nilai precision kelas olahraga yang relatif lebih kecil dibandingkan precision kelas bukan olahraga.

Dengan demikian, dari hasil yang analisis yang telah disebutkan diatas maka dapat dikatakan bahwa focused crawler akan berjalan baik pada 50 halaman maksimum karena masih memiliki tingkat relevansi yang dekat dengan link awal (*seed*). Namun secara umum terlihat bahwa naïve bayes classifier kurang dapat mengkategorikan halaman dengan baik karena terdapat penurunan akurasi jika melibatkan lebih banyak data testing. Dapat terlihat dari grafik yang semakin menurun.

5. KESIMPULAN DAN SARAN

Berdasarkan pengujian yang telah dilakukan diatas, diperoleh kesimpulan sebagai berikut:

1. Semakin banyak jumlah data training yang digunakan tidak menjamin hasil klasifikasi akan semakin baik, sesuai dengan sifat naïve bayes classification yang bersifat independen kondisional.
2. Penggunaan stopword dapat menurunkan nilai akurasi karena terdapat kata-kata umum yang tidak mewakili kelas olahraga maupun bukan kelas olahraga. Penggunaan stopword dapat meningkatkan skor relevansi karena terdapat banyak kata-kata umum yang akan dibandingkan antara dataset dan data testing.
3. Hasil focused crawler menjadi tidak baik jika bertemu dengan halaman web yang mengandung spidertrap karena halaman web akan terus mengulang dengan alamat yang berbeda padahal isi dari web tersebut sama.
4. Data training terbaik adalah Dataset100 dimana memiliki rata-rata akurasi sebesar 91,39% pada maksimum page 50. Halaman maksimum 50 menghasilkan akurasi terbaik yang dilakukan oleh naïve bayes.
5. Focused crawler bekerja dengan waktu yang relatif banyak karena harus melakukan proses

skoring dan klasifikasi dimana melibatkan ribuan term antara data training dan testing.

6. Penelusuran dengan best first search membantu dalam menelusuri halaman-halaman yang relevan terlebih dahulu. Terlihat dari hasil penelusuran bahwa web yang dikeluarkan terlebih dahulu adalah web yang berkelas aktual olahraga. Relevansi halaman ditentukan dengan hasil skoring.
7. Akurasi hasil klasifikasi halaman web dari naïve bayes terlihat dominasi menurun jika semakin banyak data testing yang terlibat.

Dengan melihat banyaknya kekurangan dari Tugas Akhir ini, maka penulis mengajukan saran agar penelitian selanjutnya dapat lebih baik. Saran yang dapat penulis berikan adalah sebagai berikut:

1. Dapat dilakukan penanganan untuk mengatasi masalah spidertrap agar halaman-halaman yang ditelusuri lebih bervariasi dan banyak karena masalah spider trap sangat berpengaruh terhadap hasil halaman yang ditelusuri.
2. Dapat dilakukan tambahan preprocessing stemming untuk data testing yang digunakan pada saat pengujian dan diharapkan dapat meningkatkan akurasi.
3. Data training yang digunakan perlu dikembangkan agar hasil klasifikasi dapat lebih baik sehingga parameter performansi lainnya juga akan meningkat.
4. Dapat digunakan algoritma penelusuran (crawling strategy) lain selain best first search, misalnya fish search, shark search atau yang lainnya.
5. Dapat melakukan penambahan variasi kategori dalam 1 mesin focused crawler sehingga pengguna dapat memilih topik yang diinginkan dengan lebih dinamis.
6. Dapat melakukan teknik klasifikasi lain yang telah dikenal, misalnya Support Vector Machine dimana menurut Gautam Pant dan Padmini Srinivasan memberikan hasil yang lebih baik.

6. REFERENSI

[1] C. Aurel, N. Deo. "*Evaluation of a Graph-based Topical Crawler"*. Available at: http://www.cs.ucf.edu/csdept/faculty/deo/icomp06-topical.pdf. Diakses tanggal 18 November 2010.

[2] Chakrabarti, S., Berg, M., and Dom, B. *"Focused crawling: A new approach to topic-specific web resource discovery"*. Computer Networks.

[3] Christopher D. Manning, Prabakhar Raghavan, and Hinrich Schutze. 2008. *"An Introduction to Information Retrieval."* Cambridge University

[4] M. Diligenti, F. M. Coetzee, S. Lawrence, C. L. Giles and M. Gori. *"Focused Crawling Using Context Graphs"*. NEC Research Institute, Princeton, NJ USA

[5] Maimunah, Siti. Kuspriyanto. "*Reinforcement Learning dalam Proses Pembelajaran Penentuan Strategi Penelusuran pada Focused Crawler".* Konferensi dan Temu Nasional Teknologi Informasi dan Komunikasi untuk Indonesia 21-23 Mei 2008, Jakarta

[6] Manger, Jason j, 1995, *"World wide web, mosaic and more"*. Mc. Graw Hill book company Europe , England.

[7] Menczer, Filippo, Gautam Pant, Padmini Srinivasan. *"Evaluating Topic-Driven Web Crawlers."*

[8] Micarelli, Alessandro and Fabio Gasparetti. *"Adaptive Focused Crawling"*. Roma Tre University.

[9] Menczer, Filippo, Gautam Pant, Padmini Srinivasan. *"Topical Web Crawler: Evaluating Adaptive Algorithm"*

[10] Partalas, I., Paliouaras, G., and Vlahavas, I., *"Reinforcement Learning with Classifier Selection for Focused Crawling"*. Available at: http://users.iit.demokritos.gr/~paliourg/papers/ECAI08_FC.pdf. Diakses tanggal 18 November 2010.

[11] *Web Crawling, Available at* web.mit.edu/aisha/Public/WebCrawling.ppt . Diakses tanggal 21 November 2010

[12] Widyantoro, D. H. "*Survey Arah Penelitian, Pengembangan dan Penerapan Penjelajah Situs Web"*. Available at: http://www.batan.go.id/sjk/eII2006/Page05/P05l.pdf . Diakses tanggal 12 Juli 2011

[13] Xuan, Wang, 2006, "*Augmenting Focused Crawling using search engine queries"*. School of Computing, National University of Singapore.