ANALISIS DAN IMPLEMENTASI MESIN FOCUSED CRAWLER UNTUK WEB OLAHRAGA DENGAN ALGORITMA BEST FIRST SEARCH

Novita Debora[1], Eko Darwiyanto[2], Erda Guslinar Perdana[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

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 topiktopik 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

1.1 Latar Belakang

Saat ini *World Wide Web* (www) bertambah sangat banyak dan berkembang begitu cepat dan menjadi salah satu sarana penyebaran informasi baik itu personal, sosial maupun komersial. Sifat dinamis (berdasarkan topologi dan konten) yang dimiliki oleh web ini menjadi salah satu faktor penunjang bagi para peneliti untuk mengembangkan *crawler* yang dapat memberikan hasil yang maksimal dalam memenuhi kebutuhan pencarian oleh pengguna internet. *Web crawler* sendiri adalah suatu program yang melakukan proses scanning ke semua halaman-halaman internet untuk dibuat indexnya dan mendukung sebuah *search engine*. Proses crawling akan memakan banyak waktu, *bandwith* dan kapasitas penyimpanan. Pertumbuhan *World Wide Web* (www) berbanding lurus dengan hal tersebut dimana membutuhkan solusi yang efisien dan *high scalable.*

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.

Untuk membedakan kelas-kelas halaman page yang relevan atau tidak dengan topik, dibutuhkan satu atau lebih classifier yang sekaligus menjadi *machine learning*. Dalam Tugas Akhir ini digunakan *Naïve Bayes Classifier*. Dalam buku *Introduction to Information Retrieval* oleh Christopher D Manning dijelaskan bahwa classifier ini memiliki kompleksitas waktu yang optimal, itu sebabnya mengapa *Naïve Bayes* merupakan metode klasifikasi teks yang paling populer disamping akurasinya yang cukup tinggi.

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*. Algoritma *Best First Search* merupakan algoritma 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. Halaman web olahraga yang akan diproses dapat berupa artikel dari orang luar yang berupa opini/pendapat, artikel khusus dari berita olahraga mengenai perlombaan, kejuaraan, liga sepakbola, dan sejenisnya, forum yang membahas mengenai dunia olahraga, serta bentuk informasi olahraga lain yang lebih serius seperti penelitian, makalah, jurnal atau sebagainya. 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.

1.2 Rumusan Masalah

Dari latar belakang permasalahan yang ada, maka masalah yang dirumuskan adalah sebagai berikut:

1. Bagaimana *Focused Crawler* mengumpulkan halaman-halaman web yang relevan dengan topik yang ditentukan?
2. Bagaimana algoritma *Best First Search* melakukan penelusuran terhadap data seed sebagai halaman penentu awal?
3. Bagaimana performansi yang dihasilkan oleh crawler dalam hal nilai akurasi, precision, recall dan F-Measure?

1.3 Batasan Masalah

Dalam pembuatan Tugas Akhir ini, untuk mengatasi permasalahan yang ada maka masalah akan dibatasi sebagai berikut:

1. Menggunakan 3 halaman uji sebagai penanda awal mesin crawler untuk melakukan penelusuran ke level berikutnya.
2. Format file yang diakses adalah HTML.
3. Topik yang dipilih untuk ditelusuri adalah web olahraga berbahasa Indonesia.
4. Data training yang digunakan diambil dari web dan dibagi menjadi 50,100, 150, 200, 250 dan 300.

1.4 Tujuan

Tujuan dari Tugas akhir ini adalah:

1. Merancang dan menganalisis suatu focused crawler dengan algoritma Best First Search dan mampu melakukan klasifikasi terhadap halaman web olahraga maupun yang bukan halaman web olahraga.
2. Mengukur performansi yang dihasilkan oleh crawler dalam hal nilai akurasi, precision recall dan F-Measure.

1.5 Metodologi Penyelesaian Masalah

Adapun metodologi penyelesaian masalah yang akan dilakukan yaitu:

1. Studi Literatur dan pengumpulan bahan

   Dalam tahap ini dilakukan pengumpulan bahan dan pembacaan khusus mengenai topik focused crawler serta algoritma Best First Search yang akan digunakan. Materi-materi yang dipelajari berupa jurnal, buku, maupun artikel dari berbagai sumber. Mencari bahan yang akan digunakan sebagai penunjang, salah satunya adalah untuk data training.

2. Analisis dan Perancangan sistem

   Setelah melakukan analisis, maka tahap selanjutnya adalah melakukan perancangan awal bagaimana nantinya focused crawler akan diimplementasikan. Perancangan sistem akan dilakukan dengan pendekatan object oriented technique berupa UML (Unified Modelling Language).

3. Implementasi Sistem

   Setelah perancangan sistem, maka akan diimplementasikan dalam bentuk baris program untuk menghasilkan mesin crawler yang telah dirancang sebelumnya.

4. Pengujian dan Evaluasi Sistem

Tahap selanjutnya adalah melakukan pengujian terhadap sistem dengan memakai data seed yang telah dipilih sebelumnya, melakukan koneksi langsung dengan internet dan melakukan pengukuran performansi. Melakukan evaluasi setelah hasil performansi didapatkan.

5. Pembuatan laporan

   Setelah sistem telah selesai diuji dan hasil pengujian sudah mencapai titik yang stabil maka dilakukan pembuatan laporan untuk mendokumentasikan semua proses yang telah dilakukan.

5. Kesimpulan Dan Saran

5.1 Kesimpulan

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

1. Klasifikasi Naïve Bayes dapat diimplementasikan pada focused crawler.
2. Semakin banyak jumlah data training yang digunakan tidak menjamin hasil klasifikasi akan semakin baik, sesuai dengan sifat naïve bayes classification yang bersifat independen kondisional.
3. Penggunaan stopword dapat menurunkan nilai akurasi karena terdapat kata-kata umum yang tidak mewakili kelas olahraga maupun bukan kelas olahraga.
4. Penggunaan stopword dapat meningkatkan skor relevansi karena terdapat banyak kata-kata umum yang akan dibandingkan antara dataset dan data testing.
5. 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.
6. Data training terbaik adalah Dataset100 dimana memiliki rata-rata akurasi sebesar 91,39% pada maksimum page 50.
7. Focused crawler bekerja dengan waktu yang relatif banyak karena harus melakukan proses skoring dan klasifikasi dimana melibatkan ribuan term antara data training dan testing.
8. Halaman maksimum 50 menghasilkan akurasi terbaik yang dilakukan oleh naïve bayes.
9. 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.
10. Akurasi hasil klasifikasi halaman web dari naïve bayes terlihat dominasi menurun jika semakin banyak data testing yang terlibat.

5.2 Saran

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.

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.