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.