Bab 1

Pendahuluan

1.1 Latar Belakang

Sejak dulu peradaban manusia tidak lepas dari suatu entitas yang bernama informasi. Segala sesuatu yang diperbuat dan dilakukan oleh manusia selalu dicatat dan disimpan, apakah itu di atas batu, daun, batang kayu, dan yang lainnya, agar dapat dibaca dan dimanfaatkan sebagai sumber ilmu bagi generasi selanjutnya. Hal ini membuktikan bahwa sudah sejak lama manusia menyadari betapa pentingnya menyimpan informasi.

Karena zaman dan budaya yang terus berganti, ilmu pengetahuan manusia semakin berkembang dan maju. Berbagai perkembangan ilmu pengetahuan tersebut disimpan dalam banyak literatur atau jurnal yang mengandung informasi penting yang sangat berguna dari masa ke masa. Oleh karena itu kebutuhan manusia terhadap media penyimpanan informasi semakin besar. Penulisan informasi pada media tertulis sudah mulai dirasakan tidak mencukupi kebutuhan penyimpanan informasi yang besar tersebut karena semakin banyak informasi maka semakin banyak media tertulis yang dibutuhkan, contohnya seperti kertas, semakin banyak kita menggunakan kertas, berarti kita membutuhkan suatu tempat penyimpanan yang besar dan dalam melakukan pencarian informasi pada media tertulis dibutuhkan ketekunan dan tenaga untuk mencari informasi pada kumpulan dokumen tertulis. Hal inilah yang mendasari manusia untuk membuat suatu media penyimpanan informasi dijital yang bisa mengurangi ketidakefisienan dan ketidakefektifan dalam penyimpanan dan perolehan informasi. Saat ini manusia sudah mengembangkan berbagai penemuan yang bersifat elektronik sebagai media penyimpanan informasi dalam bentuk dijital, seperti disket, *Hard Disk* (HD), *Compact Disc* (CD), *Digital Versatile Disc* (DVD), *Blue Ray*, *Universal*

Media Disc (UMD), dan lainnya. Dengan adanya media dijital tersebut, manusia dapat menyimpan informasi tidak hanya pada dalam bentuk tertulis, informasi dapat disimpan juga dalam bentuk video dan audio. Untuk melakukan pencarian informasi pada media dijital dapat dengan mudah dilakukan, contohnya adalah dalam melakukan pencarian dokumen dijital pada suatu perpustakaan, pencarian terhadap dokumen yang terdata tidak perlu dilakukan secara manual, tetapi dapat dilakukan oleh sistem perolehan informasi dengan menyimpan data-data identitas dari dokumen-dokumen yang tersimpan didalam perpustakaan itu.

Internet yang merupakan teknologi yang ditemukan pada akhir abad ke-20, membuat penyebaran informasi di dalam dunia maya menjadi sangat pesat. Terciptanya *World Wide Web*(WWW atau Web), format dokumen standar Web seperti *HyperText Markup Language*(HTML), dan protokol standar *HyperText Transfer Protocol*(HTTP), penyebaran dan pengaksesan terhadap informasi sudah tidak mengenal jarak dan waktu, berbagai penjuru dunia dihubungkan oleh Internet dalam dunia maya.

Saat ini Web merupakan salah satu sumber informasi yang sangat penting bagi manusia. Manusia banyak menggunakan Web untuk sarana pendidikan, mencari pengetahuan, berbagi pengalaman, memberikan informasi kepada orang lain, dan kegiatan yang biasa dilakukan oleh manusia lainnya. Semua hal itu dapat mereka lakukan dengan efektif dan efisien dengan adanya Web. Hal ini menyebabkan Web menjadi suatu sumber informasi yang kaya dan beraneka ragam isinya. Jumlah informasi yang terdapat pada Web sudah tidak terhitung lagi banyaknya, perusahaan komersial seperti Google[1] ataupun Yahoo[2] memanfaatkan keadaan ini dengan membuat suatu mesin pencari yang dapat membantu penggunanya dalam mencari informasi dalam dunia maya. Dan selain perusahaan komersial, banyak organisasi riset, yang membuat aplikasi serupa dengan menggunakan lisensi *free open source*, misalkan seperti Lucene *Search Engine*[3], Indri *Search Engine*[4], dan lain-lain.

Perolehan informasi adalah bidang ilmu yang bergerak dalam pencarian informasi di dalam suatu koleksi dokumen. Menurut Baeza-Yates [BYRN99], perolehan informasi adalah sebuah cabang dari bidang ilmu komputer yang mempelajari teknik-teknik untuk memperoleh informasi agar informasi yang diperoleh relevan dengan kueri yang dimasukkan

[1]Lihat "http://www.google.com"
[2]Lihat "http://www.yahoo.com"
[3]Lihat "http://lucene.apache.org/"
[4]Lihat "http://www.lemurproject.org/indri/"

pencari informasi. Kueri adalah suatu bahasa komputer yang digunakan untuk melakukan permintaan terhadap basis data dan sistem informasi.

Sebuah sistem perolehan informasi digunakan untuk mengurangi jumlah informasi yang berlebih, agar pencari informasi dapat lebih mudah mendapatkan informasi yang diinginkannya. Banyak universitas dan perpustakaan umum menggunakan sistem perolehan informasi untuk menyediakan akses pengguna terhadap buku, jurnal / literatur, dan dokumen lain yang terdapat di dalamnya. Dalam penelitian ini, penulis menggunakan sistem perolehan informasi Web, dimana fokus pencarian adalah informasi yang terdapat di dalam dokumen-dokumen yang terdapat di dalam Web. Sistem perolehan informasi Web yang saat ini dikenal oleh banyak orang adalah *Search Engine* atau mesin pencari. Mesin pencari adalah salah satu aplikasi sistem perolehan informasi Web, dimana mesin pencari memiliki fungsi untuk mencari informasi yang terdapat di Web. Untuk penjelasan lebih mengenai sistem perolehan informasi Web dapat dilihat pada bagian 2.2.

Dokumen yang terdapat di Web mempunyai suatu standar struktur dokumen HTML. Dokumen ini memiliki struktur yang berbeda dari dokumen biasa. Pada dokumen HTML terdapat label atau *tag* khusus untuk memberikan tanda terhadap kata atau kalimat yang nantinya dapat diinterpretasikan oleh Web *Browser*[5], hal ini yang menyebabkan perbedaan antara struktur dokumen HTML dengan struktur dokumen biasa. Selain itu, pembeda antara dokumen HTML dan biasa adalah adanya *anchor text* dan URL. Perbedaan struktur itu menyebabkan pengolahan dan pencarian informasi pada dokumen HTML berbeda dengan dokumen biasa. Hingga saat ini, format dokumen HTML telah menjadi format standar dokumen yang digunakan di lingkungan Web.

Salah satu penelitian yang dilakukan oleh Adriani dan Pandugita [AP05] adalah mengenai efek pemanfaatan struktur dokumen HTML terhadap perolehan informasi Web. Penelitian sejenis juga pernah dilakukan oleh Westerveld dkk. [WKH]. Hasil penelitian yang dihasilkan sangat bervariasi, hal ini dikarenakan koleksi yang digunakan dan aplikasi pencari yang digunakan berbeda satu sama lain. Pada penelitian yang dilakukan Adriani dan Pandugita [AP05], koleksi dokumen yang digunakan adalah koleksi dokumen EuroGOV[6]. Kesimpulan hasil penelitian yang diperoleh menunjukkan bahwa nilai evaluasi tertinggi diperoleh dengan menggunakan informasi dari badan dokumen saja. Namun, perolehan

[5]Web *Browser* adalah suatu aplikasi piranti lunak yang dapat menampilkan dan menginterpretasikan dokumen HTML terhadap pengguna

[6]Koleksi dokumen Web yang multibahasa.

sebuah dokumen relevan pada peringkat pertama dapat ditingkatkan dengan penambahan informasi dari judul dokumen. Kesimpulan yang diperoleh ini mirip dengan kesimpulan yang dibuat oleh Westerveld dkk. [WKH] yang menyatakan bahwa dengan penambahan informasi pada badan dokumen dapat meningkatkan nilai evaluasi.

Selain teknik-teknik yang telah disebutkan sebelumnya, perolehan informasi Web juga dapat ditingkatkan dengan menggunakan pendeteksian frase yang dilakukan dengan teknik penggunaan operator kedekatan kata (*term proximity*) pada kueri. Frase merupakan dua atau lebih kata yang berurutan dan memiliki arti yang berbeda jika pembentuknya dipisahkan kata demi kata. Operator kedekatan kata berdasarkan penelitian yang dilakukan oleh Okky Hendriansyah [Hen07], didapatkan bahwa dengan menggunakan teknik ini pada kueri terhadap pencarian informasi maka hasil dari perolehan informasi Web dapat ditingkatkan. Penelitian ini menggunakan berbagai kombinasi dari operator kedekatan kata dengan menggunakan Indri *Search Engine*[7].

Teknik atau metode lain dalam memperoleh informasi yang digunakan adalah umpan balik relevan semu untuk meningkatkan hasil perolehan informasi Web. Penelitian mengenai metode ini pernah dilakukan oleh Mark D. Smucker [Smu06], yang meneliti mengenai kueri yang bias akibat dari umpan balik relevan semu. Hasil dari penelitian yang dilakukan menunjukkan adanya penurunan unjuk kerja sistem dalam memperoleh informasi dengan menggunakan teknik umpan balik relevan semu jika dibandingkan dengan kueri tanpa menggunakan umpan balik relevan semu. Penggunaan umpan balik relevan semu menyebabkan bias pada kueri, sehingga kueri perluasan yang baru melebihi cakupan dari informasi yang ingin diperoleh dengan menggunakan kueri awal. Metode ini juga pernah diterapkan oleh Okky Hendriansyah [Hen07] terhadap koleksi dokumen EuroGOV, dan hasil yang didapatkan juga menurunnya kinerja dari sistem perolehan informasi.

Teknik lain yang akan diterapkan adalah pendeteksian bahasa (dapat dilihat pada bagian 2.4). Pembahasan mengenai teknik ini banyak dibicarakan pada konferensi iNEWS[8] setiap tahunnya. Tujuan utama dari iNEWS adalah untuk menemukan teknik dan mengevaluasi aplikasi yang dapat meningkatkan efektifitas dari mesin pencari. Sedangkan untuk tujuan yang lebih spesifik, iNEWS membahas beberapa topik penting yang saat ini menjadi

[7] Lihat "http://www.lemurproject.org/indri/"

[8] *Improving Non-English Web Searching* (iNEWS) adalah forum yang bertujuan untuk meningkat hasil perolehan informasi Web khususnya untuk koleksi dokumen yang tidak berbahasa Inggris

permasalahan utama dalam pencarian informasi yang tidak berbahasa Inggris, antara lain[9]:

- Mengevaluasi mesin pencari dengan menggunakan kueri yang tidak berbahasa Inggris.
- Mendefinisikan metode-metode dalam mengevaluasi efektifitas dari mesin pencari dengan menggunakan kueri yang tidak berbahasa Inggris.
- Mempelajari pola kueri yang tidak berbahasa Inggris yang diberikan pencari informasi.
- Mempelajari faktor-faktor yang mempengaruhi penggunaan mesin pencari.
- Menemukan tambahan teknik pada suatu mesin pencari yang dapat meningkatkan pencarian untuk perolehan informasi Web yang tidak berbahasa Inggris.
- Menemukan strategi pengajaran yang baik terhadap pencari informasi dalam melakukan pencarian.
- Mengidentifikasi bagaimana teknik standar pada perolehan informasi (Umpan Balik Relevan, Analisis Pranala, dll.) dapat diadaptasi dalam perolehan informasi Web yang tidak berbahasa Inggris.

Macdonald dkk. [MLO07] dalam konferensi iNEWS, melakukan penelitian yang menguji apakah teknik-teknik standar perolehan informasi Web dapat diterapkan pada perolehan informasi Web yang tidak berbahasa Inggris. Penelitian ini mengevaluasi setiap hasil perolehan berdasarkan bahasa yang terdapat pada kueri yang digunakan. Pada penelitian ini disimpulkan bahwa penggunaan teknik-teknik standar perolehan informasi Web cocok untuk diterapkan pada perolehan informasi Web yang tidak berbahasa Inggris menggunakan sistem perolehan informasi *Terrier*[10]

Pada penelitian yang dilakukan oleh Efthimiadis dkk.[EMK+08] dipelajari bagaimana tanggapan dari mesin pencari terhadap kueri yang berbahasa *Greek*. Penelitian ini bertujuan untuk meneliti apakah sebagian besar mesin pencari komersial dapat memberikan tanggapan terhadap kueri yang tidak berbahasa Inggis ataupun kueri berbahasa *Greek* untuk lebih spesifiknya. Evaluasi dilakukan terhadap mesin pencari komersial seperti Google[11]

[9]Lihat "http://rea.teimes.gr/lazarinf/ir7w/NonEnglishSearch07.pdf".
[10]Lihat "http://ir.dcs.gla.ac.uk/terrier/index.html".
[11]Lihat "http://www.google.com"

dan Yahoo[12] dan dibandingkan dengan mesin pencari lokal Yunani (*Greek*). Dari penelitian yang dilakukan didapat kesimpulan bahwa mesin pencari yang bersifat global seperti Google tidak memperdulikan karakteristik dari bahasa *Greek*, misalkan dengan memasukkan kueri *Greek* dengan ataupun tanpa aksen yang merupakan salah satu karakteristik bahasa *Greek* akan memperoleh hasil akhir yang berbeda.

Pada penelitian yang dilakukan saat ini, teknik pendeteksian bahasa akan digunakan untuk memperbaiki kinerja sistem perolehan informasi dengan melakukan pendeteksian bahasa terhadap kueri dan dokumen yang diperoleh.

Pengelompokan dokumen (*clustering*) merupakan teknik terakhir yang akan diterapkan pada penelitian Tugas Akhir ini. Tujuan dari pengelompokan dokumen adalah untuk memperbaiki tatap muka sistem terhadap pengguna agar tampilan sistem lebih baik, mengelompokkan dokumen - dokumen hasil perolehan dari sistem untuk memperbaiki *precision*[13] dan *recall*[14] pada mesin pencari, dan mengelompokkan koleksi dokumen sehingga pencarian terhadap kueri yang digunakan pada kelompok-kelompok tertentu saja [SKK]. Dalam penelitian ini, penulis mencoba melakukan pengelompokan dokumen untuk mengelompokkan dokumen-dokumen hasil perolehan dari sistem untuk meningkatkan kinerja dari sistem perolehan informasi Web. Untuk penjelasan mengenai teknik ini dapat dilihat pada bagian 2.6.

Pada penelitian ini, penulis ingin mencoba melakukan eksperimen dengan mengkombinasikan berbagai teknik tersebut, mencakup teknik pengurutan ulang analisis pranala dalam dan luar, *PageRank*, pendeteksian bahasa, dan pengelompokan dokumen berdasarkan bahasa dan topik serta teknik perolehan informasi umpan balik relevan semu, dan operator kedekatan kata terhadap koleksi dokumen multibahasa. Sehingga, dapat diketahui bagaimana kombinasi teknik terbaik yang dapat meningkatkan hasil perolehan informasi Web.

[12] Lihat "http://www.yahoo.com"
[13] *Precision* adalah parameter ketepatan sistem dalam memperoleh informasi
[14] *Recall* adalah parameter kesempurnaan atau kelengkapan sistem dalam memperoleh informasi

1.2 Rumusan Masalah

Penelitian yang dilakukan terfokus kepada permasalahan-permasalahan berikut ini:

1. Menemukan kombinasi teknik yang sebaiknya digunakan untuk meningkatkan perolehan dokumen yang relevan dan kombinasi teknik yang sebaiknya dihindari karena dapat menurunkan hasil perolehan dokumen yang relevan pada koleksi dokumen Web multibahasa.

2. Seberapa besar peningkatan atau penurunan nilai evaluasi secara umum dan berdasarkan bahasa yang terjadi dengan mengimplementasikan teknik-teknik tersebut terhadap koleksi dokumen Web multibahasa.

3. Seberapa besar peningkatan hasil perolehan informasi Web dengan mengkombinasikan teknik yang sudah diteliti oleh Hendriansyah [Hen07] yang menghasilkan nilai perolehan yang paling baik (struktur dokumen dan operator kedekatan kata) dengan teknik yang akan diterapkan pada penelitian saat ini (analisis pranala, pendeteksian bahasa, dan pengelompokan dokumen).

1.3 Tujuan dan Ruang Lingkup

Penelitian ini pada dasarnya, bertujuan untuk menemukan kombinasi teknik ataupun menemukan teknik perolehan informasi yang tepat, yang dapat meningkatkan kerelevanan atau kesesuaian informasi yang diperoleh dengan keinginan pencari informasi pada koleksi dokumen multibahasa. Koleksi dokumen Web yang digunakan dalam penelitian ini adalah koleksi dokumen Web multibahasa yang diambil dari domain pemerintahan Eropa disebut dengan EuroGOV, yang secara resmi digunakan oleh *Cross Language Evaluation Forum* (CLEF) pada tahun 2005 dan 2006.

Dalam melakukan penelitian ini, penulis menggunakan berbagai program aplikasi yang sudah tersedia untuk membantu jalannya penelitian ini (dapat dilihat pada bagian 3.2). Semua program aplikasi yang digunakan adalah aplikasi bebas dan terbuka (*free and open source software*), sehingga penulis dapat dengan bebas mengubah, menambah, dan mendistribusikan kode sumber (source code) yang tersedia.

Hasil dari penelitian ini bukanlah suatu sistem perolehan informasi ataupun sebuah mesin pencari. Penelitian ini diharapkan dapat menghasilkan suatu analisis dari teknik-

teknik perolehan informasi yang telah disebutkan sebelumnya pada bagian 1.1 dan hasil evaluasi dari teknik-teknik tersebut.

Tujuan terakhir dari penelitian ini adalah untuk mendapatkan kombinasi teknik perolehan informasi yang terbaik dalam meningkatkan nilai evaluasi terhadap kerelevanan dari perolehan informasi setinggi mungkin dengan menggunakan koleksi dokumen Web multibahasa, mengevaluasi kemampuan Indri Search Engine dalam memperoleh dokumen Web multibahasa, dan menguji apakah teknik perolehan informasi Web standar dapat diterapkan pada koleksi dokumen Web multibahasa

1.4 Studi Literatur dan Eksperimen

Dalam penulisan laporan Tugas Akhir ini, penulis menggunakan metode penelitian studi pustaka, dengan mempelajari berbagai jurnal, paper, atau literatur khusus dalam bidang perolehan informasi Web. Pustaka yang dipelajari tersebut berasal dari berbagai penelitian yang pernah dilakukan oleh orang lain mengenai teknik perolehan informasi yang akan digunakan dalam perolehan informasi.

Dengan mempelajari beberapa jurnal atau literatur tersebut, penulis dapat mengetahui teknik-teknik perolehan informasi Web dan bagaimana kinerja dan cara kerja dari teknik tersebut, sehingga penulis dapat melakukan kombinasi dari teknik-teknik yang telah diketahui itu dalam melakukan eksperimen ini.

Pada penelitian ini teknik-teknik perolehan informasi Web yang dipelajari adalah penggunaan operator kedekatan kata, penggunaan umpan balik relevan semun dan teknik-teknik pengurutan ulang dokumen hasil perolehan yaitu *PageRank*, analisis pranala dalam dan luar, pendeteksian bahasa, pengelompokan berdasarkan bahasa, dan pengelompokan berdasarkan topik secara *centroid*.

1.5 Sistematika Penulisan

Laporan Tugas Akhir ini terdiri dari lima bab, yaitu:

Bab 1 Pendahuluan membahas mengenai latar belakang penelitian, rumusan masalah, tujuan dan ruang lingkup penelitian, metodologi penelitian yang digunakan, dan sistematika penulisan.

Bab 2 Landasan Teori membahas mengenai teori dan detil dari teknik-teknik yang akan diterapkan dalam eksperimen. Selain itu pada bab ini dibahas pengertian dari perolehan informasi Web dan detil dari penelitian sebelumnya yang terkait dengan eksperimen yang akan dilakukan.

Bab 3 Eksperimen membahas mengenai alur skenario eksperimen, bagaimana eksperimen dijalankan, apa saja yang terjadi selama eksperimen berlangsung

Bab 4 Analisis Hasil Eksperimen bab ini berisi analisa penulis terhadap hasil eksperimen yang telah disebutkan pada bab sebelumnya.

Bab 5 Penutup berisi kesimpulan penulis yang dapat diperoleh dari hasil analisa eksperimen yang telah disebutkan pada bab sebelumnya. Selain itu berisi saran juga anjuran dari penulis yang sebaiknya dilakukan untuk penelitian selanjutnya pada topik yang sama.