Bab 1

Pendahuluan

Bab ini berisi penjelasan mengenai latar belakang penelitian, rumusan masalah, tujuan yang ingin dicapai dan ruang lingkup yang membatasi pelaksanaan penelitian, metodologi penelitian yang digunakan, dan sistematika penulisan makalah.

1.1 Latar Belakang

Dengan membanjirnya dan tersebarnya informasi di Internet, tanpa bantuan mesin perolehan informasi hampir tidak mungkin seorang pengguna dapat menemukan informasi yang diinginkannya di *Web*. Salah satu bentuk implementasi dari bidang ilmu perolehan informasi adalah mesin pencari (*search engine*) seperti Google[1] dan Yahoo![2].

Perolehan informasi (*information retrieval*) adalah salah satu bidang ilmu komputer yang bertujuan untuk membantu menemukan informasi tertentu di antara banyak informasi yang tersedia.

Dengan menggunakan mesin pencari, pengguna bisa mendapatkan daftar informasi di *website* yang tersebar di Internet cukup dengan memasukkan kueri (*query*) berupa sederetan kata kunci yang dianggap mewakili informasi yang diinginkan. Misalnya bila pengguna ingin mencari informasi mengenai sejarah musik zaman *renaissance* di Internet maka kueri yang bisa dimasukkan adalah "*sejarah musik klasik renaissance*" atau "*musik renaissance jerman lute harpsichord sejarah*".

Di atas kesempurnaannya, mesin pencari ini memiliki satu kelemahan. Menurut Srihari et. al. [Srihari 2000], karena kurangnya analisis linguistik terhadap teks kueri pada kebanyakan mesin pencari, maka yang sebenarnya mereka lakukan adalah aktivitas perolehan daftar dokumen dan bukan perolehan informasi. Hasil yang didapat dari mesin pencari adalah daftar dokumen atau *website*. Untuk menemukan informasi yang benar-benar diinginkan, pengguna harus memeriksa dan membaca setiap dokumen di daftar tersebut.

1 Google Search Engine (http://www.google.com)
2 Yahoo! Search Engine (http://www.yahoo.com)

Oleh karena itu kita membutuhkan mesin pencari yang bisa memberikan informasi langsung sebagai respon dari kueri yang diberikan pengguna. Bila kuerinya adalah "*Pelukis Italia yang merupakan saingan dari Pablo Picasso*" maka respon yang seharusnya diberikan cukup "*Amedeo Clemente Modigliani*", bukan daftar dokumen atau *website* yang kemungkinan berisi informasi tentang pelukis Italia tersebut. Pengguna tidak perlu lagi mencari informasi yang diinginkan dari setiap dokumen di dalam daftar, tetapi langsung mendapatkan informasi tepat seperti yang diinginkan.

Dengan motivasi ini Text REtrieval Conference[3] ke-8 (TREC, tahun 1999) , yaitu suatu konferensi internasional untuk bidang perolehan informasi, mulai merintis topik tanya jawab (*Question Answering*). Sejak saat itu banyak penelitian difokuskan secara serius untuk mengembangkan sistem yang serupa dengan mesin pencari, namun dengan kueri berupa pertanyaan dalam bahasa natural seperti "*Siapa penulis Anna Karenina?*" dan respon berupa jawaban langsung seperti "*Leo Tolstoy*". Cuplikan isi dokumen atau *website* (*snippet*) juga bisa ditampilkan untuk mendukung jawaban dan menunjukkan bagian dokumen di mana jawaban tersebut muncul.

Setelah TREC mengusung topik tanya jawab monolingual Bahasa Inggris (pertanyaan, jawaban dan koleksi dokumen dalam Bahasa Inggris) [Dang 2006], muncullah CLEF[4] (*Cross-Language Evaluation Forum*) yang memfokuskan penelitian dalam berbagai bahasa (multilingual). Di forum ini, pertanyaan bisa berada dalam Bahasa Jerman, Indonesia, atau Perancis dan koleksi dokumen yang digunakan bisa berada dalam Bahasa Inggris, Spanyol, atau Italia. Oleh karena itu isunya tidak hanya teknik penemuan jawaban yang benar tapi juga kendala bahasa.

Berturut-turut di tahun 2005 dan 2006, tim Fakultas Ilmu Komputer Universitas Indonesia telah ikut serta dalam topik tanya jawab (*Question Answering Track*) untuk subtopik bilingual Indonesia-Inggris (kueri dalam Bahasa Indonesia dan dokumen dalam Bahasa Inggris) di CLEF [Wijono 2006, Adriani 2005]. Di tahun 2007 ini penulis bersama dosen pembimbing juga meneruskan penelitian sistem tanya jawab di CLEF untuk topik yang sama.

Wijono et. al. [Wijono 2006] menggunakan pendekatan pengenalan entitas bernama atau *named entity recognition* (mengenali nama orang, tempat, atau judul buku) untuk menemukan jawaban. Jawaban yang dipilih adalah entitas bernama pada teks dokumen yang berdekatan dengan kata-kata pertanyaan. Teknik pengenalan entitas bernama

3 Text REtrieval Conference (http://trec.nist.gov).
4 Cross-Language Evaluation Forum (http://www.clef-campaign.org).

juga digunakan oleh Sarmento [Sarmento 2006], Grau et. al. [Grau 2006], dan Costa [Costa 2006].

Sementara itu Katz dan Lin [Katz 2003] menggunakan pendekatan linguistik. Mereka mengekstrak struktur tata bahasa dari setiap kalimat pertanyaan dan kalimat pada koleksi dokumen. Jawaban diperoleh dengan mencari kesamaan struktur antara dua kalimat. Misalnya kalimat pertanyaan *"What chases mouse?"* terlihat memiliki kemiripan struktur tata bahasa dengan kalimat dokumen *"Cat chases mouse"*. Maka kemudian jawaban yang diperoleh untuk pertanyaan tersebut adalah "*Cat*".

Pendekatan lain adalah menggunakan bantuan *Web* untuk menjawab pertanyaan. Neumann dan Sacaleanu [Neumann 2005] menggunakan Google untuk mengevaluasi pasangan pertanyaan dengan setiap kandidat jawabannya. Semakin banyak *hit* yang diberikan Google maka semakin besar probabilitas bahwa kandidat adalah jawaban yang benar.

Dalam penelitiannya, Laurent et. al. [Laurent 2006] menyimpulkan bahwa penggunaan sistem tanya jawab dapat menghemat waktu pencarian informasi 2 sampai 6 kali lebih cepat dibanding menggunakan mesin pencari.

Melihat berbagai penelitian tersebut, penulis ingin mencoba melakukan beberapa kombinasi teknik eksperimen mencakup pengenalan entitas bernama, penghitungan jarak kata-kata dengan *average distance weight*, menggunakan bantuan *Web*, menggunakan TF-IDF, dan pemisahan frasa kata benda menggunakan *constituency tree*. Semua teknik ini akan dijelaskan di Bab 2 Landasan Teori dan Bab 3 Eksperimen.

1.2 Rumusan Masalah

Fokus dalam penelitian ini adalah:

1. Bagaimana mengatasi kendala perbedaan bahasa secara optimal?
2. Bagaimana mendapatkan perolehan dokumen yang baik untuk dijadikan sumber pencarian jawaban?
3. Apa efek dari penggunaan informasi dari Internet untuk membantu pemilihan jawaban?

4. Bagaimana efeknya bila informasi dari Internet dikolaborasikan dengan teknik penghitungan jarak kata-kata (*average distance weight*), penggunaan TF-IDF, dan mempertimbangkan urutan dokumen?
5. Bagaimana cara mengikutsertakan teknik linguistik untuk membantu mendapatkan jawaban dari dokumen? Apa efeknya?

1.3 Tujuan dan Ruang Lingkup

Penelitian ini bertujuan untuk mencari dan mengevaluasi teknik-teknik baru yang dapat meningkatkan perolehan jawaban yang benar dari bidang sistem tanya jawab. Teknik-teknik yang berasal dari penelitian lain terutama dari tim *Question Answering* CLEF 2006 Fasilkom UI [Wijono 2006] juga diikutsertakan untuk mendukung teknik-teknik baru dari penulis.

Penelitian ini juga ditujukan dalam rangka mempersiapkan penulis untuk ikut berkontribusi di CLEF (*Cross-Language Evaluation Forum*) 2007.

Penelitian dan makalah ini juga diperuntukkan sebagai Tugas Akhir di Fakultas Ilmu Komputer Universitas Indonesia sebagai salah satu syarat kelulusan untuk program sarjana S1.

Hasil dari penelitian ini adalah sebuah sistem tanya jawab bilingual dengan masukan berupa pertanyaan dalam Bahasa Indonesia serta keluaran berupa jawaban dan cuplikan dokumen pendukung dalam Bahasa Inggris. Sistem yang dihasilkan dirancang dan dibatasi untuk menjawab pertanyaan-pertanyaan CLEF 2006.

Koleksi dokumen yang digunakan adalah dokumen koran berbahasa Inggris Glashow Herald tahun 1995 dan Los Angeles Times tahun 1994 (selanjutnya akan disebut dengan GH95 dan LA94). Koleksi dokumen ini merupakan korpus resmi dari *Question Answering Track* CLEF 2006.

1.4 Metodologi Penelitian

Penulis menggunakan dua buah metodologi dalam penelitian ini, yaitu studi literatur dan eksperimen laboratorium. Studi literatur dilakukan terhadap beberapa *paper* di bidang perolehan informasi khususnya di bidang sistem tanya jawab. Sebagian besar literatur diperoleh dari konferensi internasional seperti TREC (*Text REtrieval*

Conference) dan CLEF. Dengan melakukan studi literatur, penulis berharap dapat mengetahui teknik-teknik terbaru dari bidang sistem tanya jawab.

Setelah melakukan studi literatur, penulis mendapatkan hipotesis-hipotesis baru yang akan dibuktikan kebenarannya melalui eksperimen. Hasil evaluasi yang didapat disampaikan di bagian akhir makalah ini.

1.5 Sistematika Penulisan

Makalah ini terdiri dari enam bab yang akan dijelaskan secara singkat dalam Subbab Sistematika Penulisan ini.

Bab 1 Pendahuluan. Bagian pertama ini memaparkan mengenai latar belakang penelitian ini, rumusan masalah, tujuan dan ruang lingkup penelitian, metodologi penelitian, dan sistematika penulisan dalam makalah ini.

Bab 2 Landasan Teori. Bab ini membahas definisi-definisi dan teori-teori dalam bidang perolehan informasi, khususnya sistem tanya jawab, yang mendasari penelitian ini. Bab ini merupakan hasil rangkuman dari studi literatur yang dilakukan oleh penulis.

Bab 3 Eksperimen. Bagian ini berisi perancangan eksperimen yang dilakukan di penelitian ini. Pada bab ini juga akan diceritakan mengenai alur dan algoritma sistem tanya jawab yang dibuat oleh penulis.

Bab 4 Evaluasi dan Analisis. Di sini akan dijelaskan mengenai evaluasi yang didapat dari eksperimen yang dilakukan beserta analisisnya.

Bab 5 Kesimpulan, Saran, dan Penutup. Bab terakhir ini akan memberikan kesimpulan dari evaluasi pada bab sebelumnya. Penulis juga memberikan saran-saran mengenai hal-hal yang bisa dikembangkan untuk penelitian selanjutnya.