BAB I
PENDAHULUAN

1.1. Latar Belakang

Informasi telah menjadi bagian terpenting dari berbagai aktivitas masyarakat modern. Perkembangan teknologi Internet dan Web yang demikian pesat mengakibatkan sumber-sumber informasi menjadi semakin banyak dan beragam. Bahkan saat ini Web telah menjadi suatu kebutuhan, baik itu digunakan untuk melakukan transaksi bisnis, komunikasi, penyebaran informasi, maupun pencarian informasi.

Kehadiran mesin-mesin pencari (*search engines*) seperti Google (*www.google.com*), Yahoo (*www.yahoo.com*), Altavista (*www.altavista.com*) dan sebagainya, memberikan kemudahan untuk mencari dan menemukan informasi di Web. Namun seiring perkembangannya yang sangat pesat, saat ini terdapat milyaran dokumen Web. Peningkatan volume informasi yang sangat besar ini justru menambah kesulitan untuk menemukan, mengelola, mengakses dan memelihara informasi yang dibutuhkan. Penyebab utama timbulnya kesulitan tersebut terutama karena makna informasi yang terdapat dalam dokumen web (*web content*), hanya dapat dipahami oleh manusia namun tidak dapat dipahami oleh mesin, sehingga mesin tidak mampu menginterpretasikan informasi apa yang dibutuhkan atau dicari oleh manusia. Hal ini mengakibatkan dokumen-dokumen yang tidak relevan pun disertakan sebagai hasil pencarian (*search result*). Dan seringkali terjadi bahwa dokumen-dokumen yang relevan justru tidak terindeks

oleh mesin pencari. Sehingga campur tangan manusia untuk memilah informasi-informasi tersebut tetap dibutuhkan.

Untuk mengatasi kesulitan tersebut, dibutuhkan suatu mekanisme yang memampukan komputer memahami makna informasi yang dicari. Dengan kata lain, dibutuhkan suatu cara agar informasi dalam suatu dokumen Web dapat dibaca dan dipahami oleh mesin (*machine understandable*). Web dengan kemampuan demikian, seolah-olah memiliki kecerdasan yang sanggup memberikan jawaban yang tepat terhadap pertanyaan atau kebutuhan para penggunanya.

Semantic web (SW) yang dipelopori oleh Tim Berners-Lee, merupakan suatu cara untuk merepresentasikan *web content* dalam bentuk yang dapat dipahami dan diproses oleh mesin. Dengan kata lain, SW mengindikasikan bahwa makna data (*the meaning of data*) pada web dapat dipahami, baik oleh manusia maupun oleh komputer.

Inti dari sebuah aplikasi SW adalah pemanfaatan ontologi untuk merepresentasikan basis pengetahuan dan sumberdaya web. Ontologi menghubungkan simbol-simbol yang dipahami manusia dengan bentuknya yang dapat diproses oleh mesin, dengan demikian ontologi menjembatani kesenjangan antara manusia dan mesin.

Salah satu cara untuk menemukan informasi yang diinginkan adalah dengan memanfaatkan sistem *Question Answering* (QA). Sebuah sistem QA, menerima *query* dalam bentuk pertanyaan dengan bahasa alami, mencari jawaban pada sekumpulan dokumen atau pada basis pengetahuan dari sebuah domain,

mengekstraknya dan kemudian memformulasikan jawaban yang ringkas. Salah satu cara untuk meningkatkan kualitas sistem QA adalah dengan memanfaatkan ontologi untuk merepresentasikan basis pengetahuannya.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah dikemukakan, permasalahan yang menjadi fokus penelitian ini adalah bagaimana membangun sebuah sistem QA sederhana berbasis ontologi sebagai sebuah aplikasi SW.

1.3. Batasan Masalah

Mengingat adanya berbagai keterbatasan dan untuk menghindari kompleksitas yang mungkin timbul selama penelitian berlangsung, diberikan batasan-batasan dalam penelitian ini, yakni:

1. Domain masalah dalam penelitian ini dibatasi pada informasi film.
2. Informasi film yang dimaksud adalah artibut-atribut yang terkait dengan sebuah film, misalnya judul film, sudtradara, aktor, dan sebagainya
3. Informasi film yang digunakan bersumber pada *Internet Movie Database* (IMDB, *www.imdb.com*).
4. Model ontologi informasi film dibangun dengan menggunakan bahasa OWL (*Web Ontology Language*) dan Protégé (*ontology editor*).
5. Sistem QA sederhana dibangun untuk memroses kalimat pertanyaan dalam bahasa Indonesia.
6. Klasifikasi kalimat pertanyaan dibatasi pada pertanyaan-pertanyaan yang bersifat faktual.

7. Aplikasi SW yang dikembang berupa aplikasi pencarian menggunakan teknologi *Java Server Pages* (JSP), *Jena Ontology API* dan *SPARQL*.

1.4. Tujuan Penelitian

Penelitian ini bertujuan untuk menjawab permasalahan yang telah dikemukakan pada rumusan masalah, yakni membangun sebuah sistem QA sederhana berbasis ontologi sebagai sebuah aplikasi SW.

1.5. Manfaat Penelitian

Beberapa manfaat yang diharapkan dari hasil penelitian ini adalah sebagai berikut.

1. Bagi perkembangan ilmu, khususnya di bidang teknologi SW dan teknologi QA, penelitian ini diharapkan dapat memberikan kontribusi empiris mengenai bagaimana membangun dan mengembangkan sistem QA sederhana dan ontologi untuk sebuah domain tertentu, dan bagaimana membangun sebuah aplikasi pencarian berbasis SW.
2. Bagi para pengembang web dan pengembang sistem, penelitian ini diharapkan memberikan wawasan tentang pemanfaatan teknologi SW dan teknologi QA untuk mengembangkan sebuah sistem berbasis web.
3. Bagi para peneliti di bidang SW dan teknologi QA, penelitian ini diharapkan dapat menjadi acuan bagi penelitian lanjutan yang lebih kompleks.

1.6. Tinjauan Pustaka

Sebuah sistem QA, menerima *query* dalam bentuk pertanyaan dengan bahasa alami, mencari jawaban pada sekumpulan dokumen atau pada sebuah basis pengetahuan, mengekstraknya dan kemudian memformulasikan jawaban yang

ringkas (Moldovan & Surdeanu, 2003). Umumnya sistem QA terdiri atas tiga modul utama, yakni *question processing*, *document retrieval* dan *answer processing*. Kebanyakan sistem QA mengelompokan pertanyaan berdasarkan jenis pertanyaannya (Cooper & Ruger, 2000; Moldovan & Surdeanu, 2003; Perez-Coutino *et al*, 2004; Gunawan & Lovina, 2006; Wijono *et al*, 2006; August, 2007; Kangavari *et al*, 2008; Cheng-Lung *et al*, 2008). Jika jenis pertanyaan dapat ditentukan maka jenis jawabannya dapat ditentukan pula. Dimisalkan, jenis pertanyaannya adalah ”Siapa…” , maka jawaban yang diinginkan adalah orang atau organisasi. Jika pertanyaannya “Kapan…” jawaban yang diinginkan adalah waktu atau tanggal.

Web dengan milyaran informasi yang sangat beragam dan tak terstruktur dipandang sebagai sumber informasi yang bernilai. Walaupun saat ini tersedia banyak mesin pencari, namun mereka tidak mampu memberikan informasi yang spesifik yang diinginkan pengguna. Pemanfaatan teknologi QA pada web bertujuan untuk mengatasi masalah tersebut. Teknologi QA diharapkan dapat menjadi antarmuka yang lebih intuitif untuk memformulasikan pertanyaan dan memberikan jawaban dalam bahasa alami daripada mengembalikan sekumpulan dokumen web yang terurut berdasarkan ranking (Moldovan & Surdeanu, 2003; Perez-Coutino *et al*, 2004; McGuinness, 2004; Lopez *et al*, 2005).

Penelitian-pelitian yang terkait dengan sistem QA pada SW telah banyak dilakukan. Katz *et al* (2002) menyebutkan bahwa terdapat peluang sinerjik antara teknologi bahasa alami dan SW, yakni sebuah sistem QA yang mampu memberikan informasi yang relevan dari sebuah basis pengetahuan berbasis

ontologi dalam menanggapi *query* yang berikan oleh pengguna dalam bahasa alami.

Ide ini diwujudkan dengan mengadopsi *triple-based data model* (misalnya RDF) sebagai basis pengetahuan pada sistem QA (Katz *et al*, 2002; Lopez *et al*, 2005; Lopez, *et al*, 2006; Litkowski, 2003). Hal ini didasarkan pada pertimbangan bahwa terdapat kemungkinan untuk merepresentasikan sebuah *query* berbasis bahasa alami ke dalam bentuk *triple*, yang dalam hal ini berbentuk subyek, predikat dan obyek dari sebuah kalimat. Sementara pemodelan data dalam SW dengan menggunakan RDF (*Resource Description Framework*) juga menyatakan sebuah *statement* dalam bentuk *triple*: *resources, properties,*dan *value*.

Untuk mentransformasikan pertanyaan bahasa alami ke sebuah bentuk *query* formal digunakan metoda-metoda yang diadopsi dari teknologi *Natural Language Processing* (NLP), *Information Retrieval* (IR) dan *Information Extraction* (IE). Beberapa metoda yang sering digunakan adalah *named-entity recognition* dan *entity relation recognition*. Dalam kaitannya dengan representasi pengetahuan dalam sebuah ontologi, *named-entity* dapat dipandang sebagai sebuah *instance* atau kelas atau *value* dari sebuah properti dan *entity relation* dapat dipandang sebagai sebuah properti.

Kecenderungan penelitian-penelitan QA yang dilakukan saat ini mengarah pada *open domain* QA yang berbasis pada sejumlah besar dokumen pada web. Berbeda dengan kecenderungan tersebut, beberapa penelitian berfokus pada *restricted domain* (Lopez *et al*, 2005; Frank *et al*, 2004; Atzeni *et al*, 2004; Litkowski, 2003; Gunawan & Lovina, 2006; August, 2007; Cooper & Ruger,

2000; Kangavari *et al*, 2008). Pemilihan *restricted domain* didasarkan pada beberapa alasan, antara lain, *pertama*, eksploitasi informasi pada dokumen web sering dihadapkan pada masalah reliabilitas informasi tersebut. Dapat saja terjadi bahwa informasi yang diberikan telah kedaluwarsa atau bahkan sepenuhnya salah. *Kedua*, pemanfaatan pengetahuan formal pada *restricted domain* dapat meningkatkan keakuratan sistem QA, karena baik pertanyaan maupun jawabannya dianalisis berdasarkan basis pengetahuan tersebut. *Ketiga*, sangat dimungkinkan bahwa sebuah institusi memiliki dan mengelola basis pengetahuan yang sifatnya terbatas dan hanya dipergunakan dalam lingkup institusi tersebut.

McGuinness (2004) menyebutkan bahwa penggunaan teknologi SW dapat meningkatkan kinerja sebuah sistem QA. Hal itu dapat dilakukan dengan cara memanipulasi konten (basis pengetahuan), memanipulasi *query* atau memanipulasi jawaban. Pada umumnya sistem QA pada web, mengekstrak jawaban dari sekumpulan dokumen yang tidak terstruktur. Pada *restricted domain*, penggunaan basis pengetahuan yang terstruktur sangat dimungkinkan karena ukuran basis pengetahuannya yang cenderung lebih kecil dan stabil (Frank *et al*, 2004) dibandingkan dengan basis pengetahuan pada *open domain*. Dengan basis pengetahuan yang terstruktur (misalnya ontologi), sistem dapat menurunkan lebih banyak makna dan dapat memanfaatkan *domain* dan *range* pada *slot* untuk mengecek konsistensi informasi (McGuinness, 2004).

Sejauh ini terdapat sejumlah penelitian mengenai sistem QA yang menggunakan bahasa Indonesia (Wijono *et al*, 2006; Larasati & Manurung, 2007; August, 2007; Mahendra *et al*, 2008). Sebagai bahasa kenegaraan yang resmi,

bahasa Indonesia digunakan oleh lebih dari seratus juta orang. Berdasarkan fakta tersebut, penggunaan bahasa Indonesia sebagai bahasa alami dalam sebuah sistem QA patut dipertimbangkan.

Tabel 1.1 Penelitian-penelitian yang terkait

No.	Referensi	Teknologi	Representasi basis pengetahuan	Domain	Bahasa alami yang digunakan
1	Lopez *et al*, 2005	QA & SW	*Ontology*	*Resricted: academic life*	Inggris
2	Frank *et al*, 2005	QA & SW	*Ontology & Database*	*Restricted: nobel prize winner*	Inggris,
3	Atzeni *et al*, 2004	QA & SW	*Ontology*	*Restricted: academic life*	Italia, Denmark
4	Litkowski, 2003	QA & SW	*Ontology*	*Restricted: english news text*	Inggris
5	Katz *et al*, 2002	QA & SW	*Ontology*	*Open*	Inggris
6	Perez-Coutino *et al*, 2004	QA & SW	*Ontology*	*Open*	Spanyol
7	Lopez *et al*, 2006	QA & SW	*Ontolgy & free text*	*Open*	Inggris
8	Gunawan & Lovina, 2006	QA	*Free text*	*Restricted: World english bible*	Inggris
9	Cooper & Ruger, 2000	QA	*Free text*	*Restricted: financial news*	Inggris
10	Wijono *et al*, 2006	QA	*Free text*	*Open*	Indonesia
11	Larasati & Manurung, 2007	QA	*Free text*	*Open*	Indonesia
12	Mahendra *et al*, 2008	QA	*Free text*	*Open*	Indonesia
13	Kangavari *et al*, 2008	QA & SW	*Free text & Ontology*	*Restricted: weather information*	Inggris
14	Damljanovic *et al*, 2008	QA & SW	*Ontology*	*Open*	Inggris
15	Hoojung *et al*, 2004	QA & DBMS	*Database*	*Restricted: weather information*	Inggris
16	Cheng-Lung *et al*, 2008	QA	*Free text*	*Open*	Inggris, Cina
17	August, 2007	QA	*Free text*	*Rerstricted:* Alkitab	Indonesia

Tabel 1.1 memperlihatkan penelitian-penelitian sebelumnya yang terkait dengan penelitian yang akan dilakukan. Kesamaannya terletak pada pemanfaatan teknologi QA dan SW serta cara merepresentasikan basis pengetahuan (*ontology*), sedangkan perbedaannya terletak pada domain penelitian dan bahasa alami yang digunakan pada sistem QA. Apabila dibandingkan dengan penelitian-penelitian sistem QA yang menggunakan bahasa Indonesia, perbedaan utamanya terletak pada pemanfaatan teknologi SW, cara merepresentasikan basis pengetahuan dan domain penelitian.

1.7. Keaslian Penelitian

Penelitian-penelitian yang terkait dengan teknologi SW dan teknologi QA telah banyak dilakukan. Untuk memastikan keaslian penelitian ini, telah dilakukan serangkaian penelusuran terhadap penelitian-penelitian sebelumnya yang terkait dengan topik penelitian ini. Penelusuran secara manual dilakukan pada perpustakaan Magister Ilmu Komputer UGM dan perpustakaan S1 Ilmu Komputer UGM. Penelusuran secara *online* dilakukan melalui Google (*www.google.com*), Google Scholar (*scholar.google.com*), CiteSeer (*citeseerx.ist.psu.edu*) dan IEEE Computing Society (*www.computer.org*). Dari hasil penelusuran ditemukan beberapa penelitian yang terkait dengan topik penelitian ini (Tabel 1.1).

Penelitian ini menggunakan ontologi untuk merepresentasikan basis pengetahuan dari sistem QA berbahasa Indonesia pada sebuah domain yang terbatas (informasi film). Dengan membandingkan penelitian-penelitian

sebelumnya dan penelitian ini, dapat disimpulkan bahwa penelitian ini belum pernah dilakukan.

1.8. Metoda Penelitian

Metoda yang digunakan untuk pengembangan sistem ini adalah *waterfall model* atau *Classic Life Cycle model*. Model ini mengusulkan sebuah pendekatan yang sistematik dan sekuensial dalam mengembangkan sistem. Model ini membagi pengembangan sistem dalam lima tahap, yakni: tahap analisis, tahap desain, tahap pengkodean, tahap pengujian, dan tahap pemeliharaan (Pressman, 2001; McLeod & Schell, 2004). Sementara (Whitten *et al*, 2004) menyebutnya sebagai strategi pengembangan *model-driven*. Strategi ini terbagi atas delapan fase pengembangan, yakni: fase definisi lingkup, fase analisis masalah, fase analisis persyaratan, fase desain logis, fase analisis keputusan, fase desain dan integrasi fisik, fase konstruksi dan pengujian, dan terakhir adalah fase instalasi dan *delivery*.

Berdasarkan uraian di atas, tahapan-tahapan penelitian ini terbagi atas:

a. Tahap Definisi Lingkup Sistem

   Pada tahapan ini, seluruh masalah, kesempatan dan arahan yang mendasari pengembangan sistem ini didefinisikan. Termasuk di dalammya adalah mendefinisikan batasan sistem dan strategi pengembangan yang digunakan.

b. Tahap Analisis Persyaratan

   Pada tahapan ini, seluruh informasi yang terkait dengan pengembangan sistem dikumpulkan dan dianalisis. Informasi-informasi tersebut merupakan dasar untuk menetapkan persyaratan bisnis dari sistem yang akan dikembangkan. Penemuan fakta dilakukan dengan cara studi literatur dan *site visit*.

c. Tahap Analisis

Pada tahapan ini, sistem dimodelkan secara logis berdasarkan persyaratan-persyaratan bisnis yang telah ditentukan.

d. Tahap Desain

Pada tahapan ini persyaratan-persyaratan bisnis yang telah dimodelkan dalam tahap analisis ditransformasikan dalam spesifikasi desain fisik yang akan menjadi dasar konstruksi sistem.

e. Tahap Konstruksi dan Pengujian

Pada tahapan ini, sistem dibangun dengan menggunakan bahasa pemrograman dan tools yang telah ditentukan sebelumnya. Setelah itu akan dilakukan pengujian terhadap komponen-komponen sistem.

1.9. Sistematika Penulisan

Penulisan laporan penelitian ini terbagi dalam enam bab sebagai berikut. Bab I Pendahuluan; pada bagian ini diuraikan mengenai latar belakang topik penelitian, rumusan dan batasan masalah, tujuan dan manfaat penelitian, metoda yang digunakan dalam penelitian ini, serta tinjauan pustaka yang menjadi dasar acuan bagi penelitian ini.

Bab II Landasan teori; pada bagian ini akan dibahas teori-teori yang mendasari penelitian ini. Di antaranya teori mengenai sistem *question answering*, *semantic web*, dan ontologi.

Bab III Analisis dan Rancangan Sistem; pada bagian ini akan dibahas mengenai hasil analisis dan rancangan model sistem yang akan dibangun.

Bab IV Implementasi Sistem; pada bagian ini akan dibahas mengenai bagaimana membangun sistem *question answering* berbasis ontologi

Bab V Hasil Evaluasi dan Pembahasan; pada bagian ini akan dibahas hasil dan evaluasi dari sistem yang telah dibangun

Bab VI Kesimpulan; pada bagian ini dipaparkan simpulan dari hasil penelitian dan saran pengembangannya.