

Sistem Tanya Jawab dengan Web Semantik

A'la Syauqi

Jurusan Teknik Informatika
Fakultas Sains dan Teknologi,
Universitas Islam Maulana Malik Ibrahim Malang
Malang, Indonesia
asakuuki@ti.uin-malang.ac.id

Aeny Nurwahdah

Jurusan Teknik Informatika
Fakultas Sains dan Teknologi,
Universitas Islam Maulana Malik Ibrahim Malang
Malang, Indonesia
aku.ainee@gmail.com

Abstrak—Saat ini internet merupakan sarana favorit kebanyakan masyarakat dalam memperoleh informasi. Untuk mendapatkan informasi yang diinginkan mereka dengan mudah memasukkan kata kunci ke dalam layanan search engine yang telah tersedia. Namun informasi yang dimunculkan seringkali tidak sesuai dengan harapan. Pengguna harus memilih links (yang menyambung ke halaman lain) yang isinya terkadang tidak sesuai dengan kebutuhan. Penelitian ini bertujuan untuk menyelesaikan permasalahan tersebut dengan membuat sistem tanya jawab berbasis teknologi web semantik agar pengguna memperoleh informasi/jawaban yang sesuai dengan kebutuhan. Untuk penggunaan diperlukan masukan berupa kalimat tanya yang kemudian diproses melalui empat tahapan untuk penemuan jawaban, ialah: teks processing, pencarian resource, eksekusi query sparql, dan penampilan jawaban. Dari pengujian ditunjukkan bahwa aplikasi mampu mengembalikan jawaban dengan baik. Akurasi ketepatan jawaban yang didapat mencapai 83,81%.

Kata kunci—aplikasi tanya jawab, web semantik, stemming, natural language processing

I. PENDAHULUAN

Internet telah menjadi media yang mempermudah bagi setiap orang untuk mengakses sumber pengetahuan dan kebudayaan. Hal ini dimungkinkan karena adanya teknologi web yang semakin berkembang dari tahun ke tahun. Saat ini untuk mewujudkan website, baik organisasi atau perorangan bukan merupakan pekerjaan yang rumit lagi. Dengan begitu penyampaian informasi, ide, gagasan adalah hal yang mudah dilakukan. Demikian juga yang terjadi di Indonesia. Indonesia merupakan salah satu negara terbesar pengguna internet. Berdasarkan survei yang diselenggarakan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), jumlah pengguna internet di Indonesia mencapai 71,19 juta [1]. Dari jumlah tersebut, prosentase tertinggi pengguna internet di Indonesia dari sektor industri adalah pemanfaatan email 95,75%, disusul kemudian pencarian informasi atau berita 78,49%. Sedangkan di sektor konsumen, pemanfaatan internet untuk media sosial 88% dan sebagai sarana pencarian informasi atau berita 68% [2]. Dari data tersebut menunjukkan bahwa pemanfaatan internet untuk pencarian informasi atau berita menempati posisi yang penting.

Proses penemuan kembali informasi yang berguna melalui internet bisa saja menjadi sulit karena banyaknya dokumen web yang ada. Untuk memilah dokumen web yang sesuai dengan kebutuhan, pengguna dapat memanfaatkan layanan search engine yang telah ada, misalnya: Google, Yahoo, Bing, dan sebagainya. Tetapi hasil yang disajikannya kurang spesifik, sebagai contoh untuk mendapatkan informasi yang diinginkan pengguna harus memilih links (yang menyambung ke halaman lain) yang terkadang tidak sesuai dengan kebutuhan.

Dari masalah tersebut maka perlu dilakukan pembaharuan dalam teknik pencarian. Salah satunya yakni dengan membuat Sistem tanya jawab. Sistem tanya jawab merupakan suatu sistem yang mengizinkan pengguna menyatakan kebutuhan informasinya dalam pertanyaan, kemudian mengembalikan kutipan teks singkat sebagai jawaban. Sistem tanya jawab berbasis web yang dibangun bertujuan agar pengguna mendapat jawaban yang sesuai dengan kebutuhan.

II. PENELITIAN TERKAIT

Penelitian lain tentang teknologi semantic web diterapkan pada sistem rekam medis elektronik yang terintegrasi dengan sistem tanaman obat dengan metode sharing data [3]. Data rekam medis yang tersimpan dalam database, dapat diakses dalam konteks semantik dengan melakukan mapping dalam format RDF (resource description framework). Proses query data dilakukan dengan menggunakan query SPARQL. Dari hasil ujicoba menunjukkan data dapat diakses melalui RDF map tanpa harus mengakses database secara langsung sehingga data dapat dikelola user sesuai dengan kebutuhan.

Sistem tanya jawab pernah dibuat menggunakan metode yang navigasi pengguna agar diperoleh jawaban yang benar. Metode ini berusaha untuk memfokuskan dalam proses pencarian jawaban jika pertanyaan yang diajukan oleh pengguna mengandung ketidakpastian. Hal ini dilakukan dengan mengkerucutkan topik yang diinginkan pengguna berdasarkan kata kunci dari dokumen-dokumen yang diperoleh. Kemudian pencarian dilakukan kembali berdasarkan kata kunci tersebut [4].

Teknik lain yang pernah digunakan dalam sistem jawab dengan mengotomasi kategorisasi yang interaktif. Metode ini

terdiri dari empat tahap: konstruksi space feature, identifikasi dan pembobotan word berdasarkan topik, pemetaan semantik, dan perhitungan nilai similaritas [5].

Penelitian tentang Sistem tanya jawab (Question Answering System) yang lain pernah dikerjakan dengan metode rule-based. Dari penelitian dihasilkan aplikasi desktop dengan menggunakan dokumen Terjemahan Al Qur'an Surah Al Baqarah yang tersimpan dalam database relasional. Proses recall jawaban dimulai dengan memecah (parsing) suatu dokumen menjadi kalimat-kalimat. Kalimat-kalimat tersebut dipecah dan di-stem menjadi token. Begitu pula dengan kalimat pertanyaan pada query dipecah dan di-stem menjadi token. Token dari setiap kalimat dokumen maupun kalimat query diproses dalam rule sesuai dengan tipe pertanyaannya. Proses di dalam rule memberikan nilai untuk masing-masing kalimat dokumen. Kalimat yang memiliki nilai tertinggi akan dikembalikan sebagai jawaban. Akurasi rata-rata rule terhadap kueri dalam penelitian adalah 85.69%, sedangkan akurasi rata-rata rule terhadap kueri pengguna umum adalah 53.14% [6].

Question Answering System lainnya dikerjakan dengan menggunakan dokumen terjemah Juz Amma sebagai sumber pengetahuan yang tersimpan dalam database. Dalam penelitian ini digunakan metode wordmatch scoring dan rule based scoring. Penggunaan wordmatch scoring bertujuan untuk pemberian skor berdasarkan kesesuaian kata pada pertanyaan dan tipe jawaban. Skor digunakan untuk menentukan kandidat jawaban berdasarkan pertanyaan yang diajukan oleh user. Skor juga digunakan untuk perangkingan jawaban/berdasarkan hasil pengukuran relevansi pada sistem presicion tertinggi adalah jawaban dari kata tanya siapa dan mengapa. Recall tertinggi adalah jawaban dari kata tanya mengapa. Accuracy tertinggi adalah jawaban dari kata tanya mengapa [7].

Penelitian tentang penggalian informasi menggunakan Wikipedia sebagai basis pengetahuan juga telah dikerjakan. Penelitian ini berfokus pada ekstraksi dan membuat penggunaan konsep, relasi, fakta dan deskripsi yang ditemukan di Wikipedia. Proses pengerjaannya dibagi menjadi empat kategori, yaitu : penggunaan Wikipedia ke dalam Natural Language Processing, pemanfaatan Wikipedia sebagai fasilitas Information Retrieval dan Information Extraction, serta menggunakan Wikipedia sebagai sumber dari pembangunan ontologi [8].

Pada penelitian ini dibuat sistem tanya jawab dengan penerapan semantik web dimana terdapat empat tahap dalam pemrosesan: teks processing, pencarian resource, eksekusi query SPARQL dan retrieve jawaban dari DBpedia.

III. WEB SEMANTIK

Semantic Web adalah perkembangan generasi berikutnya atau yang biasa disebut sebagai evolusi dari WWW (World Wide Web), yang dicetuskan pada tahun 2002. Semantic Web adalah Web yang mencakup dokumen, atau bagian dari dokumen, menggambarkan hubungan eksplisit antara hal dan berisi informasi semantik ditujukan untuk pemrosesan otomatis oleh mesin (komputer) [9]. W3C (World Wide Web Consortium) mendefinisikan format metadata tersebut adalah Resource Description Framework (RDF). Tiap unit dari RDF adalah 3 komposisi, yaitu subject, predicate, dan object [10].

Subject dan object adalah entitas yang ditunjukkan oleh teks. Sedangkan predicate adalah komposisi yang menerangkan sudut pandang dari subject yang dijelaskan object. Hal yang paling menarik dari RDF yaitu object dapat menjadi subject yang nantinya diterangkan oleh object yang lainnya. Sehingga object atau masukan dapat diterangkan secara jelas dan detail, serta sesuai dengan keinginan pengguna yang memberikan masukan.

Dalam mencapai tujuannya dibutuhkan pemberian meaning kedalam masing-masing content (sebagai atribut) yang akan digunakan oleh teknologi web semantic kedalam beberapa layer [11].

A. Ontology

Ontology dapat didefinisikan sebagai suatu cara untuk mendeskripsikan arti dan relasi dari istilah-istilah. Ontology merupakan suatu teori. Deskripsi tersebut berisi classes terkadang juga disebut concept dan instances. Deskripsi ini dapat membantu sistem komputer dalam menggunakan istilah-istilah tersebut cara yang lebih mudah.

B. SPARQL

SPARQL Protocol dan RDF Query Language (SPARQL) adalah sebuah *protocol* dan bahasa query untuk *semantic web resources* [12]. SPARQL merupakan salah satu bahasa query yang digunakan untuk melakukan query pada RDF (*Resource Description Framework*). Sama halnya dengan query SQL, SPARQL digunakan untuk *retrieve* data-data yang diinginkan. SQL digunakan untuk query pada data dari database yang terdiri dari satu atau beberapa tabel sedangkan SPARQL ditujukan untuk query terhadap data pada RDF yang berupa triple. Query pada SQL dapat dilakukan cukup dengan mengetahui nama tabel dan atribut atau kolom tabel. Sedangkan untuk melakukan query SPARQL pada suatu file RDF setidaknya harus mengetahui resource, property dan value atau triple pada RDF tersebut [13]. Untuk mengetahui triple pada RDF bisa dengan cara melakukan bentuk query DESCRIBE. Sebuah query yang menggunakan SPARQL dapat terdiri atas *triple patterns*, konjungsi (or) dan disjungsi (and). Kegunaan SPARQL dalam Sistem tanya jawab yang akan dibangun adalah untuk proses *retrieve* jawaban dari sumber pengetahuan DBpedia.

C. DBpedia

DBpedia merupakan sebuah komunitas yang bergerak untuk mengekstrak informasi terstruktur dari Wikipedia dan menyediakan informasi tersebut dalam sebuah web [14]. Untuk mengakses pengetahuan (knowledge) DBpedia, digunakan URI (Uniform Resource Identifier) resource dari entitas dalam bentuk SPARQL [14]. Apabila URI resource diakses menggunakan Semantic Web agents, maka hasil yang ditampilkan adalah RDF descriptions. Sedangkan apabila diakses menggunakan Web Browser akan ditampilkan informasi dari entitas yang diakses dalam tampilan HTML sederhana.

D. Stemming

Stemming merupakan suatu proses yang terdapat dalam sistem IR (Information Retrieval) yang mentransformasikan kata-kata yang terdapat dalam suatu kalimat ke kata-kata akarnya (root word) dengan menggunakan aturan-aturan tertentu. Stemming untuk bahasa yang satu berbeda algoritma stemming dengan bahasa lainnya. Sebagai contoh bahasa Inggris memiliki morfologi yang berbeda dengan bahasa Indonesia sehingga algoritma stemming untuk kedua bahasa tersebut juga berbeda. Pada umumnya kata dasar pada bahasa Indonesia terdiri dari kombinasi :

Prefix 1 + Prefix 2 + Kata dasar + Sufiks 3 + Sufiks 2 + Sufiks 1

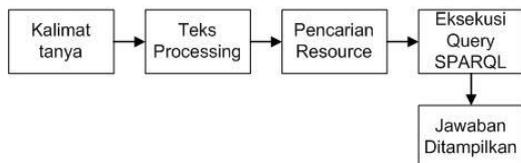
Pada teks berbahasa Inggris, proses yang diperlukan hanya proses menghilangkan sufiks. Sedangkan proses stemming pada teks berbahasa Indonesia lebih kompleks karena terdapat imbuhan yang harus dibuang untuk mendapat kata dasar (root word) dari sebuah kata.

Dari penelitian banyak algoritma yang telah ditemukan untuk keperluan stemming dengan keunggulan dan kelemahan masing-masing. Untuk bahasa Indonesia terdapat beberapa algoritma yang populer ialah algoritma Nazief and Adriani, algoritma Arifin and Setiono, algoritma Vega, serta algoritma Ahmad, Yusoff, and Sembok. Sistem tanya jawab yang akan dibangun menggunakan algoritma stemming Nazief and Adriani karena dinilai paling efektif [15].

IV. METODE

A. Deskripsi Sistem

Pada penelitian ini tujuan pembangunan sistem adalah membuat aplikasi berbasis web untuk penggalian informasi dari internet dalam bentuk tanya jawab. Aplikasi ini dibangun dengan teknologi semantic web dengan sumber pengetahuan yang berasal dari DBpedia Indonesia. Sistem dari aplikasi yang dibangun terdiri dari beberapa tahap seperti ditunjukkan pada diagram blok gambar 1



Gambar 1. Desain Sistem

Berikut penjelasan secara rinci tahap-tahap dari sistem yang dibangun:

1) Tahap pertama, pengguna menginputkan pertanyaan. Selanjutnya kalimat tanya mengalami text preprocessing yang bertujuan untuk menghilangkan stop word dan penemuan kembali akar kata/kata dasar dari setiap kata dalam kalimat pertanyaan.

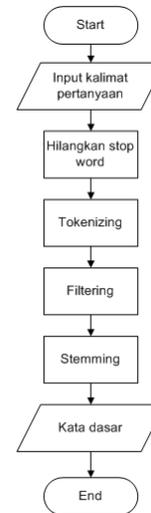
2) Kemudian penentuan resource. Kata dasar yang diperoleh dari tahap sebelumnya digunakan dalam proses ini. Proses ini dilakukan dengan cara crawling website wikipedia, untuk memperoleh link url dokumen yang sesuai dengan resource Dbpedia.

3) Tahap selanjutnya ialah proses ekstraksi teks dari resource Dbpedia dengan query SPARQL.

4) Terakhir, informasi dari hasil ekstraksi query SPARQL ditampilkan kepada pengguna sebagai jawaban.

B. Text Preprocessing

Teks preprocessing adalah tahap pengolahan kalimat yang dimaksudkan untuk menghilangkan stop word dan memperoleh akar kata/kata dasar dari setiap kata dalam kalimat tersebut. Tahap ini terdiri dari tiga bagian yaitu tokenizing, filtering, dan stemming seperti ditunjukkan dalam gambar 2



Gambar 2. Flowchart Text Preprocessing

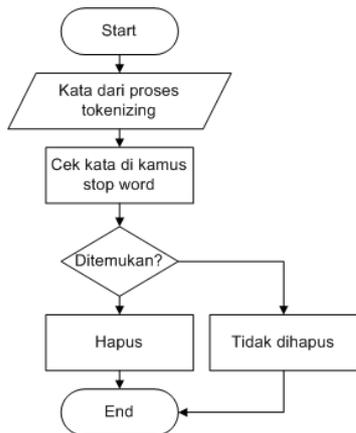
Uraian dari masing-masing tahap tersebut adalah sebagai berikut:

1) Tokenizing bertujuan untuk memecah kalimat menjadi satuan kata atau token. Algoritma dari tokenizing sebagaimana ditunjukkan gambar 3. Proses ini dimulai dengan penghilangan karakter atau simbol yang tidak diperlukan. Kemudian dilanjutkan dengan casefolding untuk merubah semua huruf teks menjadi huruf kecil. Proses ini diakhiri dengan pemecahan kalimat berdasarkan spasi sehingga diperoleh kata-kata penyusun kalimat.



Gambar 3. Flowchart Tokenizing

2) Filtering digunakan untuk menghilangkan kata-kata yang dianggap tidak penting. Algoritma filtering ditunjukkan pada gambar 4. Proses filtering dilakukan dengan cara pencocokan kata dengan kamus stop word. Apabila kata cocok dengan kata yang ada dalam kamus stop word maka akan dihapus.



Gambar 4. Flowchart Filtering

3) Stemming digunakan untuk menghilangkan imbuhan pada kata sehingga diperoleh akar kata/kata dasarnya. Pada aplikasi yang dibangun diterapkan algoritma stemming Nazief & Adriani karena algoritma ini memiliki prosentase precision yang besar [16].

C. Pencarian Resource

Pencarian resource bertujuan untuk penemuan URI resource pada DBpedia yang diperlukan pada saat proses eksekusi SPARQL. Agar diperoleh URI resource yang sesuai, digunakan google custom search API untuk penelusuran artikel pada website Wikipedia. Gambar 5 menunjukkan tampilan dokumen wikipedia tentang internet.

Internet

Dari Wikipedia bahasa Indonesia, ensiklopedia bebas

Artikel ini sudah memiliki daftar referensi, bacaan terkait atau pranala luar, tapi sumbernya masih belum jelas karena tak memiliki kutipan pada kalimat (catatan kaki). Mohon tingkatkan kualitas artikel ini dengan memasukkan rujukan yang lebih mendetail bila perlu.

Internet (kependekan dari *interconnection-networking*) adalah seluruh jaringan komputer yang saling terhubung menggunakan standar sistem global *Transmission Control Protocol/Internet Protocol Suite* (TCP/IP) sebagai protokol pertukaran paket (*packet switching communication protocol*) untuk melayani miliaran pengguna di seluruh dunia. Rangkaian internet yang terbesar dinamakan **Internet**. Cara menghubungkan rangkaian dengan kaidah ini dinamakan *internetworking* ("antarmaringan").



Gambar 5. Dokumen Wikipedia Tentang Internet

Penelusuran dilakukan menggunakan keyword yang diperoleh dari teks preprocessing. Dari proses ini diperoleh beberapa link URL website wikipedia yang sesuai dengan URI resource DBpedia. Tampilan dokumen DBpedia seperti pada gambar 6. Berikut contoh kesesuaian link URL dan URI resource: untuk artikel wikipedia dengan URL:

<http://id.wikipedia.org/wiki/Internet>

maka memiliki URI resource:

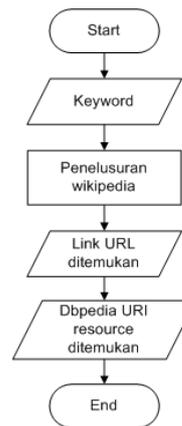
DBpedia <http://id.dbpedia.org/page/Internet>.

Sedangkan algoritma pencarian resource ditunjukkan pada gambar 7.

The screenshot shows a DBpedia entry for 'Internet'. It includes metadata like 'Content-Length: 115723' and 'About: Internet'. The main content is a definition of Internet in Indonesian: 'Internet (kependekan dari interconnection-networking) adalah seluruh jaringan komputer yang saling terhubung menggunakan standar sistem global Transmission Control Protocol/Internet Protocol Suite (TCP/IP) sebagai protokol pertukaran paket (packet switching communication protocol) untuk melayani miliaran pengguna di seluruh dunia. Rangkaian internet yang terbesar dinamakan Internet. Cara menghubungkan rangkaian dengan kaidah ini dinamakan internetworking ("antarmaringan").'

Gambar 6. Dokumen DBpedia Tentang Internet

Untuk mendapatkan resource dilakukan beberapa tahap yaitu input keyword pada sistem, keyword tersebut kemudian diproses menggunakan Google Custom Search Engine. Dari proses tersebut diperoleh indeks link yang selanjutnya diekstrak menggunakan format JSON. Dari hasil data berformat JSON tersebut kemudian dilakukan proses pemotongan url dengan mengambil bagian path file atau resource-nya. Sehingga didapatkan suatu resource yang dibutuhkan untuk proses selanjutnya.



Gambar 7. Flowchart Pencarian Resource

D. Query SPARQL

Resource yang telah diperoleh dari proses sebelumnya, kemudian digunakan sebagai variabel URI dalam query SPARQL. Dalam istilah SPARQL, resource merupakan bagian dari URI resource yang digunakan sebagai penunjuk yang membawa informasi dalam suatu web. Gambar 8 menunjukkan contoh query SPARQL yang digunakan dalam aplikasi:

```

$query=
"select ?jawaban
where {

  dbpedia-id:Internet dbpedia-owl:abstract
  ?jawaban
}";
  
```

Gambar 8. Implementasi Resource pada Query SPARQL

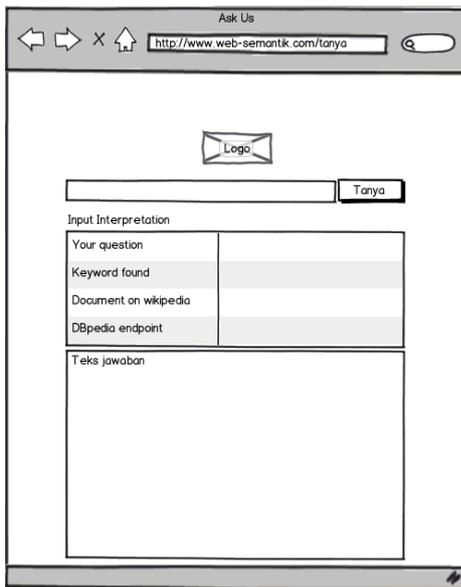
Dari eksekusi SPARQL tersebut diperoleh hasil berupa teks seperti yang ditunjukkan pada gambar 9

jawaban
"Internet (kependekan dari interconnection-networking) adalah seluruh jaringan komputer yang saling terhubung menggunakan standar sistem global Transmission Control Protocol/Internet Protocol Suite (TCP/IP) sebagai protokol pertukaran paket (packet switching communication protocol) untuk melayani miliaran pengguna di seluruh dunia. Rangkaian internet yang terbesar dinamakan Internet. Cara menghubungkan rangkaian dengan kaidah ini dinamakan internetworking ("antarjaringan")."@id

Gambar 9. Hasil query SPARQL

E. Desain Interface

Tampilan Aplikasi Tanya Jawab yang dibangun didesain sesederhana mungkin dengan tujuan untuk memudahkan user dalam mengakses. Aplikasi tanya jawab berbasis web terdiri dari satu halaman yang memuat logo, textfield untuk mengajukan pertanyaan, button "tanya" untuk penemuan jawaban, tabel untuk menampilkan interpretasi pertanyaan, dan textarea untuk menampilkan hasil jawaban. Desain interface ditunjukkan pada gambar 10.



Gambar 10. Desain Interface

Pada tabel dimuat hasil interpretasi yang terdiri dari empat bagian yaitu :

- 1) Pertanyaan yang telah diajukan
- 2) Keyword yang ditemukan
- 3) Link document pada Wikipedia
- 4) Resource pada DBpedia

Sedangkan pada text area dibagi menjadi dua bagian utama yaitu:

- 1) Document extraction memuat hasil eksekusi query SPARQL yang merupakan jawaban untuk pertanyaan yang telah diajukan.
- 2) Source berisi links Wikipedia yang dirasa sesuai dengan jawaban yang dimunculkan.

HASIL DAN PEMBAHASAN

A. Pengujian

Skenario pengujian dibuat untuk memperoleh hasil tingkat keakuratan aplikasi dalam menyajikan jawaban terhadap

pertanyaan yang diajukan oleh pemakai. Objek yang digunakan untuk pengujian adalah kalimat pertanyaan Bahasa Indonesia. Kalimat pertanyaan yang diproses harus berbahasa Indonesia baku. Kata tanya yang diajukan adalah dengan kata tanya 5W + 1H, ialah:

1) Apa (what)

Kata tanya "apa" digunakan untuk awalan kalimat tanya dengan jawaban kata benda.

2) Di mana (where)

Kata tanya "di mana" digunakan untuk awalan kalimat tanya dengan jawaban tempat atau lokasi.

3) Siapa (who)

Kata tanya "siapa" digunakan sebagai awalan kalimat apabila kalimat tanya memerlukan jawaban berupa orang, group, kelompok, dan sebagainya.

4) Kapan (when)

Kata tanya "kapan" digunakan sebagai awalan pada saat kalimat tanya memerlukan jawaban berupa waktu.

5) Mengapa (why)

Kata tanya "mengapa" digunakan sebagai awalan pada kalimat tanya dimana jawaban yang diperlukan berupa sebab.

6) Bagaimana (how)

Kata tanya "bagaimana" digunakan sebagai awalan pada kalimat tanya dengan jawaban berupa deskripsi.

7) Berapa

Kata tanya "berapa" digunakan pada awal kalimat tanya yang memerlukan jawaban berupa bilangan atau kuantitas suatu objek.

Klasifikasi jenis pertanyaan yang digunakan dalam pengujian ditunjukkan pada tabel 1. Aplikasi ini dapat memproses pertanyaan yang memerlukan jawaban berupa bilangan, seperti pertanyaan tentang jarak, tinggi, usia, dan sebagainya. Akan tetapi, aplikasi tidak dapat memproses pertanyaan yang berbentuk perhitungan dan aritmatika.

TABEL 1. KLASIFIKASI JENIS PERTANYAAN

Kata Tanya	Tipe Jawaban	Contoh Pertanyaan
Apa	Benda	Apa yang dimaksud dengan <i>smartphone</i> ?
Dimana	Lokasi	Dimana letak makam Bung Karno?
Siapa	Orang	Siapa nama walikota Surabaya?
Kapan	Waktu	Kapan Indonesia dinyatakan merdeka?
Mengapa	Alasan	Mengapa Jepang menyerah kepada Belanda tahun 1945?
Bagaimana	Deskripsi	Bagaimana proses metamorfosis berlangsung?
Berapa	Integer (jarak, tinggi, usia dll)	Berapa tinggi tugu Monas?

Hasil pengujian yang telah dilakukan terhadap aplikasi seperti ditunjukkan pada tabel 2, tabel 3, tabel 4, tabel 5, tabel 6, tabel 7, dan tabel 8. Masing-masing tabel menunjukkan jenis kalimat tanya beserta ketepatan jawaban yang dihasilkan.

TABEL 2. TABEL UJI COBA PERTANYAAN TENTANG BENDA

NO	PERTANYAAN	JAWABAN	
		TEPAT	TIDAK
1	Apa yang dimaksud dengan <i>smartphone</i> ?	√	
2	Apa warna bendera negara Jerman?	√	
3	Apakah pengertian dari <i>hardware</i> ?		√
4	Apa nama maskot kota Surabaya?	√	
5	Apa arti semboyan Bhineka Tunggal Ika?	√	
6	Apa nama tarian khas Bali?		√
7	Apa makanan khas Yogyakarta?	√	
8	Apakah judul lagu kebangsaan Indonesia?		√
9	Apa teori yang ditemukan oleh Albert Einstein?		√
10	Apa nama planet yang paling dekat dengan matahari?	√	
11	Apa nama planet yang memiliki cincin?	√	
12	Apa nama kota yang disebut kota hujan?	√	
13	Apa makanan khas Lamongan?	√	
14	Apa nama gunung di kota Batu?	√	
15	Apa nama jembatan di kota Palembang?	√	

TABEL 3. UJI COBA PERTANYAAN TENTANG LOKASI/TEMPAT

NO	PERTANYAAN	JAWABAN	
		TEPAT	TIDAK
1	Dimana letak sungai Musi?	√	
2	Dimana letak tugu Monas?	√	
3	Dimana letak museum Radya Pustaka?	√	
4	Dimana letak makam Bung Karno?	√	
5	Dimanakah letak Jembatan Ampera?	√	
6	Dimanakah Pangeran Diponegoro dimakamkan?	√	
7	Dimana letak kerajaan Majapahit?	√	
8	Dimana letak Gunung Bromo?	√	
9	Dimana tugu pahlawan berada?	√	
10	Dimana tempat ibadah umat muslim?	√	
11	Dimana letak jembatan suramadu?	√	
12	Dimana letak pulau Raja Ampat?	√	
13	Dimana lokasi Institut Teknologi Telkom?	√	
14	Dimana lokasi Taman Safari Indonesia 2?	√	
15	Dimana kota asal bahasa Osing?	√	

TABEL 4. UJI COBA PERTANYAAN TENTANG ORANG

NO	PERTANYAAN	JAWABAN	
		TEPAT	TIDAK
1	Siapa nama gubernur Jawa Timur?	√	
2	Siapa nama istri Bung Karno?	√	
3	Siapa pembuat sosial media Facebook?	√	
4	Siapakah nama walikota Surabaya?	√	
5	Siapakah CEO Apple Inc?	√	
6	Siapakah nama kiper Arema?	√	
7	Siapa pencipta lagu Indonesia Raya?	√	
8	Siapa pencetus teori gravitasi?		√
9	Siapa pencetus teori relativitas?	√	
10	Siapakah CEO Microsoft saat ini?	√	
11	Siapakah proklamator kemerdekaan Indonesia?	√	
12	Siapa vokalis Ungu band?	√	
13	Siapa penemu listrik?		√
14	Siapa nama istri SBY?	√	
15	Siapa nama gubernur Bank Indonesia?	√	

TABEL 5. UJI COBA PERTANYAAN TENTANG WAKTU

NO	PERTANYAAN	JAWABAN	
		TEPAT	TIDAK
1	Kapan Indonesia dinyatakan merdeka?	√	
2	Kapan Gus Dur mulai menjabat sebagai presiden?	√	
3	Kapan Dude Harlino lahir?	√	
4	Kapan Nagita Slavina memulai karir?	√	
5	Kapan tanggal lahir Bunga Citra Lestari?	√	
6	Kapan sholat Idul Adha berlangsung?	√	
7	Kapan perayaan natal berlangsung?	√	
8	Kapan masa jabatan Jokowi sebagai Gubernur DKI Jakarta?	√	
9	Kapan Hari Kebangkitan Nasional diperingati?		√
10	Kapan Apollo 11 tiba di Bulan?	√	
11	Kapan Soeharto berhenti menjadi presiden?	√	
12	Kapan PKI dibubarkan?	√	
13	Kapan TNI dibentuk?	√	
14	Kapan terjadinya perjanjian hudaibiah?	√	
15	Kapan terjadinya perang badar?	√	

TABEL 6. UJI COBA PERTANYAAN TENTANG SEBAB

NO	PERTANYAAN	JAWABAN	
		TEPAT	TIDAK
1	Mengapa Jepang menyerah terhadap Belanda tahun 1945?	√	
2	Mengapa terjadi perang sampit?	√	
3	Mengapa Idul Adha disebut juga hari raya haji?	√	
4	Mengapa Jogja dijuluki sebagai daerah istimewa?		√
5	Mengapa terjadi gerhana bulan?	√	
6	Mengapa terjadi gerakan Aceh merdeka?	√	
7	Mengapa Bogor disebut kota hujan?	√	
8	Mengapa Bali disebut pulau Dewata?		√
9	Mengapa Pontianak disebut kota Khatulistiwa?	√	
10	Mengapa terjadi krisis moneter tahun 1998?	√	
11	Mengapa Korea berpisah menjadi 2 wilayah?	√	
12	Mengapa Malin Kundang dikutuk menjadi batu?	√	
13	Mengapa Surabaya disebut kota Pahlawan?	√	
14	Mengapa Edelweiss disebut bunga abadi?		√
15	Mengapa Timor Leste berpisah dari Indonesia?	√	

TABEL 7. UJI COBA PERTANYAAN TENTANG DESKRIPSI

NO	PERTANYAAN	JAWABAN	
		TEPAT	TIDAK
1	Bagaimana proses metamorfosis berlangsung?	√	
2	Bagaimana proses membuat tempe?	√	
3	Bagaimanakah proses terbentuknya pelangi?		√
4	Bagaimana cara menanam jagung?		√
5	Bagaimana cara menginstal ulang Windows?		√
6	Bagaimanakah cara merawat Kucing Anggora?		√
7	Bagaimana cara mengukur presisi?	√	
8	Bagaimana aturan permainan gobak sodor?		√
9	Bagaimana cara mengukur akurasi?	√	
10	Bagaimana aturan permainan catur?	√	
11	Bagaimana proses terjadinya hujan?	√	
12	Bagaimana proses terjadinya fotosintesis?	√	
13	Bagaimana cara mencangkok tumbuhan?	√	
14	Bagaimana cara menanam padi?		√
15	Bagaimana proses terjadinya gerhana matahari?	√	

TABEL 8. UJI COBA PERTANYAAN TENTANG NILAI

NO	PERTANYAAN	JAWABAN	
		TEPAT	TIDAK
1	Berapa panjang jalan tol Surabaya-Gempol?	√	
2	Berapa tinggi tugu Monas?	√	
3	Berapa banyak pulau di Indonesia?	√	
4	Berapa luas Kebun Raya Purwodadi?	√	
5	Berapa banyak suku yang ada di Indonesia?	√	
6	Berapa tinggi gunung Bromo?	√	
7	Berapa luas wilayah Taman Nasional Bromo Tengger Semeru?	√	
8	Berapa diameter bunga Raflesia?	√	
9	Berapa jarak Bumi dengan Matahari?	√	
10	Berapa lama proses rotasi Bumi?	√	
11	Berapa jarak dari surabaya ke jakarta?	√	
12	Berapa personel JKT48?	√	
13	Berapa tekanan darah normal manusia?	√	
14	Berapa warna yang ada pada pelangi?		√
15	Berapa jumlah episode Running Man?	√	

Setiap jawaban yang dihasilkan dari masing-masing pertanyaan selanjutnya dilakukan pengukuran performansi. Pengukuran yang digunakan adalah *recall*, *precision*, dan *accuracy*. Pengukuran *recall* digunakan untuk mengetahui kemampuan sistem untuk memanggil dokumen yang relevan. Sedangkan *precision* digunakan untuk mengetahui kemampuan sistem untuk tidak memanggil dokumen yang tidak relevan.

Dari hasil pengujian diperoleh beberapa kemungkinan ialah:

- 1) True Positive (TP) yaitu jawaban yang dihasilkan sistem benar
- 2) False Positive (FP) yaitu jawaban yang dihasilkan salah atau sistem tidak menghasilkan jawaban
- 3) True Negative (TN) yaitu pertanyaan yang diajukan tidak sesuai dengan ketentuan dan sistem tidak menghasilkan jawaban
- 4) False Negative (FN) yaitu pertanyaan yang diajukan tidak sesuai dengan ketentuan tetapi sistem menghasilkan jawaban

Berikut rumus yang digunakan untuk menghitung *precision* dan *recall* beserta *accuracy* pada sistem :

$$Precision = \frac{TP}{TP + FP} \times 100\%$$

$$Recall = \frac{TP}{TP + FN} \times 100\%$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100\%$$

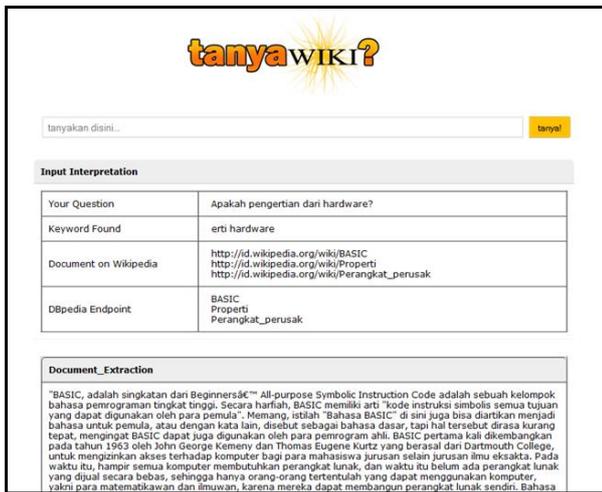
Dari rumus *precision*, *recall*, dan *accuracy* maka dihasilkan masing-masing nilainya yang ditunjukkan pada tabel 9

TABEL 9. TABEL PERBANDINGAN HASIL UJI COBA
PERTANYAAN KESELURUHAN

Precision	Recall	Accuracy
83,81%	100%	83,81%

B. Analisa

Berdasarkan uji coba yang telah dilakukan, diketahui bahwa aplikasi tanya jawab yang telah dibangun mampu menemukan kembali jawaban dengan baik untuk setiap pertanyaan yang diajukan. Keakuratan aplikasi tanya jawab dengan adanya Algoritma *stemming* Nazief & Adriani pada pemrosesan pertanyaan mencapai 83,81%.



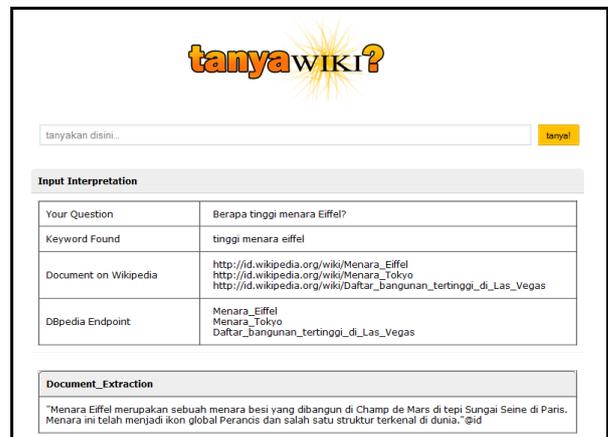
Gambar 11. Overstem yang Menyebabkan Kerancuan dalam Menemukan Keyword

Hasil analisa terhadap data uji coba diketahui bahwa akurasi jawaban dipengaruhi oleh hasil *stemming* yang menjadi *keyword* dalam pencarian *resource*. Jenis kalimat tanya tidak mempengaruhi hasil jawaban, karena kata tanya yang digunakan tidak diproses pada sistem. Sistem memproses *keyword* hasil *text-preprocessing*. Apabila terdapat *overstem* pada hasil penemuan *keyword*, maka terjadi ketidaksesuaian *resource* yang ditemukan sehingga jawaban yang ditampilkan menjadi rancu. Gambar 11 menunjukkan contoh hasil *overstem* yang menyebabkan *keyword* tidak sesuai dengan pertanyaan yang diajukan. Pada proses *stemming* tersebut, seharusnya *keyword* yang didapatkan adalah kata "arti hardware" atau "pengertian hardware". Namun, karena terjadi *overstem* maka *keyword* beserta jawaban yang diperoleh tidak sesuai dengan yang diharapkan.



Gambar 12. Contoh Pencarian Tinggi Monas

Selain terjadinya *overstem*, penyebab tidak ditemukannya jawaban adalah data tidak ditemukan pada dokumen yang ada pada DBpedia. Hal ini dikarenakan tidak semua artikel yang terdapat pada DBpedia mencakup segala pengetahuan yang dibutuhkan. Sebagai contohnya, ketika *user* bertanya mengenai tinggi Monas seperti gambar 12 aplikasi memberikan jawaban yang sesuai. Namun ketika *user* bertanya mengenai tinggi menara Eiffel, aplikasi hanya memberikan jawaban berupa deskripsi singkat mengenai menara Eiffel dan tidak memberikan jawaban yang diharapkan seperti gambar 13. Hal ini dikarenakan pada dokumen menara Eiffel yang terdapat pada DBpedia tidak terdapat data mengenai tinggi menara tersebut.



Gambar 13. Contoh Pencarian Tinggi Menara Eiffel

KESIMPULAN

Kesimpulan yang diperoleh dari penelitian ini adalah sebagai berikut :

- 1) Aplikasi Tanya Jawab dibangun dengan menerapkan semantic web pada sistem. Semantic web menjadi proses inti penggalian jawaban. Untuk mendapatkan jawaban, dilakukan pemrosesan pertanyaan yakni diawali dengan tokenizing, filtering dan stemming. Metode stemming yang digunakan adalah Algoritma Nazief & Adriani. Dari hasil pemrosesan pertanyaan tersebut diperoleh keyword untuk melakukan

penemuan resource. Resource digunakan sebagai URI resource atau petunjuk letak dokumen pada DBpedia. Setelah resource ditemukan, dokumen DBpedia tersebut kemudian diekstrak dengan menggunakan bahasa query SPARQL. Hasil ekstrak dokumen selanjutnya ditampilkan sebagai jawaban atas pertanyaan yang telah diajukan.

2) Akurasi jawaban yang diperoleh dengan adanya pemrosesan pertanyaan menggunakan algoritma stemming Nazief & Adriani mencapai 83,81%. Stemming merupakan bagian penting dalam suatu sistem temu balik karena hasil stemming mempengaruhi relevansi jawaban yang dihasilkan oleh sistem.

REFERENSI

- [1] Sinaga, R., 2014. APJII: pengguna internet di Indonesia terus meningkat. [Online] Available at: <http://www.antaranews.com/berita/414167/apjii-pengguna-internet-di-indonesia-terus-meningkat> [Accessed 24 March 2014]
- [2] APJII, 2014. Profil Terkini Internet Industri Indonesia. [Online] Available at: <http://www.apjii.or.id/v2/read/content/info-terkini/213/press-release-profil-terkini-internet-industri-ind.html> [Accessed 23 March 2014].
- [3] Ana, N. & Syauqi, A., 2012. Implementasi Teknologi Semantic Web Pada Dokumentasi. In Seminar Nasional Ilmu Pengetahuan Teknik. Bandung, 2012. LIPI.
- [4] Fukumoto, J., Aburai, N., & Yamanishi, R., 2013. Interactive Document Expansion for Answer Extraction of Question Answering System. *Procedia - Procedia Computer Science*, 22, 991–1000. doi:10.1016/j.procs.2013.09.184
- [5] Song, W., Wenyan, L., Gu, N., Quan, X., & Hao, T., 2011. Automatic categorization of questions for user-interactive question answering. *Information Processing*, 47(2), 147–156. doi:10.1016/j.ipm.2010.03.002
- [6] Anggraeny, Meinar Dwi. 2008. Skripsi : Implementasi Question Answering dengan Metode Rule-Based pada Terjemahan Al Qur'an Surat Al Baqarah. Bandung : IPB.
- [7] Lutfi, Citra Rosiana. 2012. Skripsi : Question Answering System pada terjemah juz Amma menggunakan metode Rule Based. Malang : UIN Maulana Malik Ibrahim.
- [8] Medelyan, O., Milne, D., Legg, C., & Witten, I. H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9), 716–754. Artificial Intelligence; Computation and Language; Information Retrieval. doi:10.1016/j.ijhcs.2009.05.004
- [9] W3C, W. W. W. C. (1999). Web Architecture: Describing and Exchanging Data. Retrieved 26 Februari, 2012, from <http://www.w3.org/1999/04/WebData#References>.
- [10] W3C, W. W. W. C. (2002). Resource Description Framework (RDF): Concepts and Abstract Syntax. Retrieved 26 Februari, 2012, from <http://www.w3.org/TR/2002/WD-rdf-concepts-20021108/>.
- [11] Berners-Lee, T. (2000). Semantic Web - XML2000, slide 10. Retrieved 26 Februari, 2012, from <http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html>.
- [12] Clark, K. G., Grant, K., & Torres, E., 2008. SPARQL Protocol for RDF. W3C Recommendation. Retrieved from <http://www.w3.org/TR/rdf-sparql-protocol/>
- [13] Arenas, M., & Pérez, J., 2011. Querying semantic web data with SPARQL. In Proceedings of the 30th symposium on Principles of database systems of data - PODS '11 (p. 305). New York, New York, USA: ACM Press. doi:10.1145/1989284.1989312
- [14] Mendes, P. (2011). DBpedia. dbpedia.org. Retrieved from <http://dbpedia.org/About>
- [15] Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S. M. M., & Williams, H. E., 2007. Stemming Indonesian. *ACM Transactions on Asian Language Information Processing*, 6(4), 1–33. doi:10.1145/1316457.1316459
- [16] Agusta, Ledy. 2009. Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief & Adriani untuk Stemming Dokumen Teks Bahasa Indonesia. Bali: Konferensi Nasional Sistem dan Informatika