Klasifikasi Dokumen Berita Teks Bahasa Indonesia menggunakan Ontologi

Herny Februariyanti dan Eri Zuliarso
Fakultas Teknologi Informasi, Universitas Stikubank
email: hernyfeb@gmail.com, ezuliarso@yahoo.com

Abstrak

Salah satu cara yang paling berhasil untuk mengorganisasikan informasi dalam jumlah banyak dan dapat dipahami oleh para pencari informasi adalah dengan melakukan klasifikasi dokumen berdasarkan topiknya. Kebutuhan akan dokumen pembelajaran untuk melakukan klasifikasi dokumen merupakan salah satu permasalahan yang sering muncul dalam topik klasifikasi dokumen. Permasalahan yang timbul menjadi semakin rumit dengan adanya fakta bahwa jumlah simpanan data berita menjadi sangat besar dan tidak terorganisir. Oleh karena itu, diperlukan suatu strategi pengelompokan otomatis dokumen-dokumen berita tersebut.

Klasifikasi merupakan salah satu metode dalam data mining yang bertujuan untuk mendefinisikan kelas dari sebuah objek yang belum diketahui kelasnya. Pada klasifikasi terlebih dahulu akan dilakukan proses training dan testing. Pada proses tersebut akan digunakan dataset yang telah diketahui kelas objeknya.

Pada penelitian ini akan dibangun aplikasi Klasifikasi Berita Menggunakan Ontologi. Obyek penelitian dari penelitian ini adalah artikel berita berbahasa Indonesia dari situs http://www.google.com.

Dengan adanya klasifikasi dokumen maka hasil download berita dari situs http://www.google.com dapat lebih terstruktur sehingga untuk mendapatkan informasi lebih cepat dan relevan sesuai dengan yang diinginkan.

Kata kunci: klasifikasi, bencana, ontology

PENDAHULUAN

Dengan adanya teknologi internet mempermudah bagi siapapun untuk mendapatkan informasi ataupun berita-berita yang diinginkan. Informasi yang diperoleh dari internet dapat berupa dokumen teks seperti dokumen berita, suara, video, maupun objek multimedia lainnya. Informasi tersebut dapat diakses melalui halaman web. Salah satu cara yang paling berhasil untuk mengorganisasikan informasi dalam jumlah banyak dan dapat dipahami oleh para pencari informasi adalah dengan melakukan klasifikasi dokumen berdasarkan topiknya. Dengan kemudahan untuk mendapatkan informasi dan banyaknya informasi yang didapatkan dari internet menyebabkan permasalahan baru. Kebutuhan akan dokumen pembelajaran untuk melakukan klasifikasi dokumen merupakan salah satu permasalahan yang sering muncul dalam topik klasifikasi dokumen. Permasalahan yang timbul adalah bagaimana mendapatkan informasi ataupun berita-berita yang sesuai dengan yang kita inginkan dalam waktu yang singkat. Permasalahan yang timbul menjadi semakin rumit dengan adanya fakta bahwa jumlah simpanan data berita menjadi sangat besar dan tidak terorganisir. Oleh karena itu, diperlukan suatu strategi pengelompokan otomatis dokumen-dokumen berita tersebut.

Untuk mempermudah pencarian informasi yang sesuai dengan yang kita inginkan dan sesuai dengan waktunya, maka pengklasifikasifikasian dokumen akan membantu bagaimana mendapatkan informasi, sehingga mempermudahan pengolahan dan penggunaannya sesuai kebutuhan dan tujuan yang ingin dicapai.

Klasifikasi merupakan salah satu metode dalam data mining yang bertujuan untuk mendefinisikan kelas dari sebuah objek yang belum diketahui kelasnya. Pada klasifikasi terlebih dahulu akan dilakukan proses training dan testing. Pada proses tersebut akan digunakan dataset yang telah diketahui kelas objeknya.

Permasalahan lain yang muncul adalah seberapa banyak dokumen pembelajaran yang dibutuhkan agar klasifikasi dokumen memberikan akurasi yang maksimal. Apabila jumlah dokumen pembelajaran yang digunakan terlalu sedikit, maka tidak akan menghasilkan tingkat akurasi yang maksimal. Permasalahan dokumen pembelajaran untuk melakukan klasifikasi dokumen ini dapat diatasi dengan pendekatan baru yang tidak memerlukan dokumen pembelajaran. Pendekatan ini dikenal dengan nama pendekatan ontologi.

METODE PENELITIAN

Obyek penelitian dari penelitian ini adalah artikel berita berbahasa Indonesia dari situs http://www.google.com.

Data Yang diperlukan

Merupakan data yang mendukung dalam penelitian ini meliputi data primer dan data sekunder.

1. Data primer

Data yang diperoleh langsung dari situs http://www.google.com

2. Data Sekunder

Data yang diperoleh dengan membaca dan mempelajari referensi mengenai klasifikasi dokumen, ontologi, komponen ontologi, teks mining, klasikasi dokumen menggunakan ontologi.

3. Teknik Pengumpulan Data

Pengumpulan data mempunyai tujuan mendapatkan materi – materi yang mempunyai keterkaitan dengan topik penelitian. Pengumpulan data dimaksudkan agar mendapatkan bahan-bahan yang relevan, akurat dan reliable. Maka teknik pengumpulan data yang dilakukan dalam penelitian ini adalah dengan metode Observasi, Studi Pustaka dan Metode pengembagan dengan menggunakan model prototyping.

PEROLEHAN INFORMASI

Istilah perolehan informasi memiliki pengertian yang sangat luas, sehingga banyak pakar mendefinisikan istilah perolehan informasi dari berbagai sudut pandang. Baeza- Yates dan rekannya (Baeza-Yates, 1999) memberikan definisi tentang perolehan informasi, yaitu "sebuah cabang ilmu dari ilmu komputer yang mempelajari teknik-teknik untuk memperoleh informasi (bukan data) yang relevan berdasarkan kueri yang dimasukkan oleh pencari informasi".

Perolehan informasi berbeda dengan perolehan data. Perolehan informasi merujuk pada representasi, penyimpanan, pengorganisasian sampai ke pengaksesan informasi (Baeza-Yates, 1999) Representasi dan pengorganisasian informasi harus memudahkan pencari informasi dalam mengakses informasi yang terdapat pada koleksi. Sementara itu, perolehan data memiliki lingkup yang lebih sempit. Perolehan data, dalam konteks sistem perolehan informasi, merujuk pada cara untuk menentukan atau mencocokkan antara kata-kata yang terkandung di sebuah dokumen dengan kata-kata yang digunakan seseorang dalam melakukan pencarian informasi (Baeza-Yates, 1999)

Informasi dapat berupa teks, gambar, suara, video dan obyek multimedia lainnya. Informasi dalam bentuk teks merupakan fokus utama dalam penelitian ini. Informasi merupakan sesuatu yang tidak dapat didefinisikan secara tepat. Informasi berhubungan dengan bahasa alami yang biasanya tidak terstruktur dan secara semantik dapat memiliki makna ganda atau ambigu. Masalah yang muncul kemudian adalah bagaimana caranya untuk memperoleh informasi yang relevan di antara informasi lain dalam suatu koleksi dokumen. Hal inilah yang kemudian mendorong banyaknya penelitian tentang perolehan informasi khususnya informasi dalam bentuk teks.

KLASIFIKASI DOKUMEN

Klasifikasi dokumen adalah bidang penelitian dalam perolehan informasi yang mengembangkan metode untuk menentukan atau mengkategorikan suatu dokumen ke dalam satu atau lebih kelompok yang telah dikenal sebelumnya secara otomatis berdasarkan isi dokumen (Tenenboim, L., dkk., 2008). Klasifikasi dokumen bertujuan untuk mengelompokkan dokumen yang tidak terstruktur ke dalam kelompok-kelompok yang menggambarkan isi dari dokumen. Dokumen dapat berupa dokumen teks seperti artikel berita. Pada bagian ini membahas tentang penelitian dalam bidang klasifikasi artikel berita berbahasa Indonesia. Penelitian yang dilakukan oleh Yudi Wibisono yaitu klasifikasi berita berbahasa Indonesia menggunakan Naïve Bayes classifier (Wibisono, Y., 2005). Dokumen teks dibagi menjadi dua bagian yaitu dokumen pembelajaran dan dokumen pengujian. Hasil eksperimen penelitian ini adalah metode Naïve Bayes classifier memiliki akurasi yang tinggi yaitu 89,47%. Nilai akurasi tetap tinggi terutama jika dokumen pembelajaran yang digunakan besar (lebih besar atau sama dengan 400). Kesimpulan yang diperoleh dari penelitian ini adalah metode Naïve Bayes classifier terbukti dapat digunakan secara efektif untuk mengklasifikasikan berita secara otomatis. Penelitian yang dilakukan oleh Slyvia Susanto yaitu pengklasifikasian dokumen berita berbahasa Indonesia dengan menggunakan Naïve Bayes classifier (stemming atau non-stemming) (Susanto, S., 2006). Eksperimen yang dilakukan dalam penelitian ini dengan menggunakan stemming dan non- stemming.

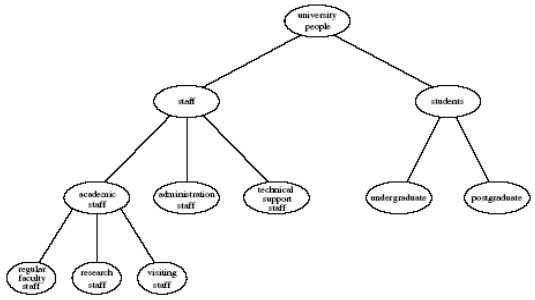
KNOWLEDGE ENGINEERING

Pendekatan *knowledge engineering* disebut *rule base* karena pendekatan ini memanfaatkan keahlian manusia (*human expert*) untuk membuat aturan-aturan (*rules*) secara manual melalui proses pemahaman pada sebuah domain penelitian (Milton, N., 2003). Dalam penelitian ini, *human expert* atau pakar dituntut untuk bisa memahami sebuah domain yang digunakan dalam pemodelan ontologi untuk klasifikasi dokumen secara otomatis. Dengan pendekatan *rule base* ini, nilai akurasi klasifikasi dokumen menggunakan ontologi sangat tergantung dari pakar yang membuat aturan-aturan yang digunakan dalam klasifikasi dokumen. Kelebihan dari pendekatan *rule base* adalah dengan menggunakan keahlian manusia untuk mencapai nilai akurasi klasifikasi dokumen yang tinggi. Pendekatan ini tidak terlalu sulit untuk dilakukan selama terdapat pakar yang memahami domain yang digunakan untuk klasifikasi dokumen dengan baik. Akan tetapi, hal inilah yang menjadi kelemahan *rule base*, yaitu metode klasifikasi dokumen menggunakan ontologi sangat bergantung pada adanya pakar. Selain itu, pendekatan ini memiliki kekurangan lain, yaitu membutuhkan waktu yang panjang dan biaya yang tinggi. Biaya yang tinggi ini disebabkan kebutuhan terhadap sumber daya manusia yang banyak terlebih jika domain yang digunakan untuk klasifikasi dokumen memiliki ruang lingkup yang sangat besar. Metode klasifikasi dokumen dengan menggunakan pendekatan *rule base* juga akan mengalami masalah *adaptability*, yaitu ketika pakar yang membuat aturan-aturan dalam sistem sudah tidak ada sehingga pakar yang baru sulit untuk melakukan penyesuaian jika ingin melakukan perubahan pada domain. Oleh karena itu, pendekatan *rule base* cocok untuk digunakan jika terdapat pakar yang memahami domain penelitian.

ONTOLOGI

Istilah ontologi berasal dari filsafat. Dalam kontek ini, ontologi digunakan sebagai subbidang dari filsafat, yang mempelajari sifat alami dari kebradaan, cabang dari metafisik yang berkaitan dengan identifikasi, atau secara umum, jenis-jenis benda yang secara actual ada, dan bagaimana memaparkannya. Sebagai contoh, observasi yang dilakukan pada objek tertentu yang mengelompokkan menjadi kelas-kelas abstrak berdasarkan pada sifat-sifat bersama merupakan komitmen ontologi secara tipe. Namun demikian, dalam beberapa tahun belakangan,ontologi menjadi kata yangdiambil oleh ilmu komputer dan diberikan sebuah arti teknis khusus yang sedikit berbeda dari aslinya.

Secara umum, sebuah ontologi memaparkan secara formal sebuah domain topik pembicaraan. Sebuah ontologi terdiri dari sebuah daftar istilah terbatas dan hubungan diantara istilah-istilah ini. *Istilah* menandakan pentingnya *konsep (kelas dari objek)* dari suatu domain. Sebagai contoh, dalam suatu universitas, ada anggota staff, mahasiswa, matakuliah dan disiplin ilmu adalah konsep yang penting. Hubungan (*relationship*) mencakup hirarki dari kelas-kelas. Sebuah hirarki menspesifikasikan sebuah kelas menjadi subkelas C_ yang lain jika setiap objek dalam C juga termasuk dalam C_. Sebagai contoh, diperlihatkan hirarki sebuah domain universitas.



Gambar 1. Contoh Ontologi Domain Universitas

Dalam kontek web, ontologi menyediakan pemahaman bersama dari suatu domain. Pemahaman bersama dibutuhkan untuk mengatasi perbedaan terminology. Sebagai contoh, di universitas A membuka program studi Ilmu Komputer, sedang di universitas B dinamakan Teknik Informatika. Beberapa perbedaan dapat diatasi dengan memetakan terminologi tertentu ke ontologi bersama atau dengan mendefinisikan pemetaan langsung diantara ontologi.

Ontologi sangat berdaya guna untuk organisasi dan navigasi situs web. Banyak situs web saat ini menyajikan sisi sebelah kiri halaman tingkat tertinggi dari hirarki konsep suatu istilah.

Komponen Ontologi

Ontologi memilki beberapa komponen yang dapat menjelaskan ontologi tersebut, diantaranya (Coral Calero, dkk., 2006):

1. Konsep (*Concept*)

   Digunakan dalam pemahaman yang luas. Sebuah konsep dapat sesuatu yang dikatakan, sehingga dapat pula merupakan penjelasan dari tugas, fungsi, aksi, strategi, dan sebagainya. *Concept* juga dikenal sebagai *classes*, *object* dan *categories.*

2. Relasi (*relation*)

   Merupakan representasi sebuah tipe dari interaksi antara konsep dari sebuah domain. Secara formal dapat didefinisikan sebagai subset dari sebuah produk dari n set, R:C1 x C2 x…xCn. Sebagai contoh dari relasi binary termasuk *subclass-of* dan *connected-to.*

3. Fungsi (*functions*)

   Adalah sebuah relasi khusus dimana elemen ke-n dari relasi adalah unik untuk elemen ke n-1. F:C1 x C2 x …Cn-1 – > Cn, contohnya adalah Mother-of.

4. Aksioma (*axioms*)

   Digunakan untuk memodelkan sebuah *sentence* yang selalu benar.

5. Instances

   Digunakan untuk merepresentasikan elemen.

Ada beberapa langkah yang diperlukan untuk mengembangkan ontologi, yaitu : (Noy., N.F., 2001)

1. Tahap penentuan domain

   Tahap ini merupakan tahap awal proses digitalisasi pengetahuan yang dilakukan dengan menjawab beberapa pertanyaan seperti apa yang menjadi domain ontologi.

2. Tahap penggunaan ulang ontologi

   Dalam tahap ini, kita melakukan pengecekan apakah ontologi yang sudah ada dapat digunakan kembali atau kita perlu mengembangkan ontologi dari awal. Apabila kita menggunakan ontologi yang sudah ada kemudian kita melakukan perbaikan dan memperluas ontologi yang sudah ada, maka kita dapat lebih menghemat waktu dari pada mengembangkan ontologi dari awal.

3. Tahap penyebutan istilah-istilah pada ontologi

   Tahap ini menentukan semua istilah penting yang digunakan untuk membuat pernyataan atau menjelaskan hal yang mirip atau sama. Contoh *class* "*wines*" berhubungan dengan istilah *wine*, anggur, lokasi, warna, bentuk, rasa dan kadar gula.

4. Tahap pendefinisian *class* dan hierarki *class*

   Tahap ini membuat definisi dari *class* dalam bentuk hierarki dan kemudian menguraikan *property* dari *class*. Hierarki *class* merepresentasikan sebuah relasi "*is-a*" (sebuah *class* A adalah *subclass* dari B jika setiap *instance* dari B adalah juga sebuah *instance* di A).

5. Tahap pendefinisian *property*

   Tahap ini mendefinisikan *property* dari masing-masing *class* yang ada di ontologi.

6. Tahap pendefinisian *facets*

   Tahap ini mendefinisikan *facets* dari setiap *property* yang ada di *class* pada ontologi.

7. Tahap mendefinisikan *instances*

   Tahap ini mendefinisikan sebuah *instance* dari suatu *class* meliputi pemilihan *class*, pembuatan individu *instance* dari *class*, dan pengisian nilai *property*.

Klasifikasi Dokumen Menggunakan Ontologi

Proses pengklasifikasian artikel berita berbahasa Indonesia terdiri atas dua langkah, yaitu: proses penemuan kosa kata kunci dalam dokumen dan kosa kata tersebut dipetakan ke sebuah node dalam konsep hierarki (ontologi). Proses pemetaan dilakukan setelah melakukan proses persiapan dokumen dan pembobotan kata.Proses persiapan dokumen teks meliputi proses case folding, tokenisasi, pembuangan stopwords dan pemotongan imbuhan (Baeza-Yates, R., 1999).

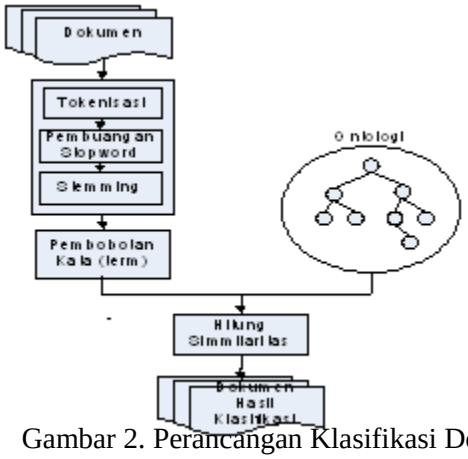
Tujuan dari proses persiapan dokumen teks adalah untuk menghilangkan karakter-karakter selain huruf, menyeragamkan kata dan mengurangi volume kosa kata. Proses pembobotan kata adalah proses memberikan nilai atau bobot ke sebuah kata berdasarkan kemunculannya pada suatu dokumen teks (Baeza-Yates, R.,1999). Proses persiapan dokumen teks dalam penelitian ini menghasilkan kumpulan kata atau term yang kemudian direpresentasikan dalam sebuah terms vector. Terms vector dari suatu dokumen teks d adalah tuple bobot semua term pada d. Nilai bobot sebuah term menyatakan tingkat kepentingan term tersebut dalam merepresentasikan dokumen teks. Pada penelitian ini, proses pembobotan kata menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF).

Tujuan penghitungan IDF adalah untuk mencari kata-kata yang benar-benar merepresentasikan suatu dokumen teks padasuatu koleksi. Metode pembobotan kata yang digunakan dalam penelitian ini adalah metode TF-IDF. Metode ini digunakan karena metode ini paling baik dalam perolehan informasi (Khodra, L.M., & Wibisono, Y., 2005). Rumus TF-IDF dapat dilihat pada Persamaan (1) (Salton, M., 1983).

$$tfidf(i,j) = tf(i,j) \times \log\left(\frac{N}{df(j)}\right)$$

dimana tf(i,j) adalah frekuensi kemunculan term j pada dokumen teks $d_i \in D^*$, dimana i =1,2,3,...,N, df(j) adalah frekuensi dokumen yang mengandung term j dari semua koleksi dokumen,dan N adalah jumlah seluruh dokumen yang ada di koleksi dokumen.

Perancangan klasifikasi dokumen teks dengan menggunakan ontologi dapat dilihat pada gambar 2



Gambar 2. Perancangan Klasifikasi Dokumen Menggunakan Ontologi

TEXT MINING

Text mining dapat diartikan sebagai penemuan informasi yang baru dan tidak diketahui sebelumnya oleh komputer, dengan secara otomatis mengekstrak informasi dari sumber-sumber yang berbeda. Kunci dari proses ini adalah menggabungkan informasi yang berhasil diekstraksi dari berbagai sumber (Hearst, 2003). Sedangkan menurut (Harlian Milkha, 2006) text mining memiliki definisi menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen, dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen.

Dengan *text mining* tugas-tugas yang berhubungan dengan penganalisaan teks dengan jumlah yang besar, penemuan pola serta penggalian informasi yang mungkin berguna dari suatu teks dapat dilakukan. Sebagai bentuk aplikasi dari *text mining*, sistem klasifikasi berita menggunakan berita sebagai sumber informasi dan informasi klasifikasi sebagai informasi yang akan diekstrak dari sumber informasi. Informasi klasifikasi dapat berbentuk angka-angka probabilitas, set aturan atau bentuk lainnya.

Tahapan *Text Mining*

1. Text Preprocessing

Tahapan awal dari *text mining* adalah *text preprocessing* yang bertujuan untuk mempersiapkan teks menjadi data yang akan mengalami pengolahan pada tahapan berikutnya. Beberapa contoh tindakan yang dapat dilakukan pada tahap ini, mulai dari tindakan yang bersifat kompleks seperti *partofspeech* (pos) *tagging*, *parse tree*, hingga tindakan yang bersifat sederhana seperti proses parsing sederhana terhadap teks, yaitu memecah suatu kalimat menjadi sekumpulan kata. Selain itu pada tahapan ini biasanya juga dilakukan *case folding*, yaitu pengubahan karakter huruf menjadi huruf kecil.

2. Text Transformation (feature generation)

Pada tahap ini hasil yang diperoleh dari tahap *text preprocessing* akan melalui proses tranformasi. Adapun proses transformasi ini dilakukan dengan mengurangi jumlah kata-kata yang ada dengan penghilangan *stopword* dan juga dengan mengubah kata-kata ke dalam bentuk dasarnya (*stemming*). *Stopword* adalah kata-kata yang bukan merupakan ciri (kata unik) dari suatu dokumen seperti kata sambung, kata kepunyaan. Memperhitungkan *stopword* pada transformasi teks akan membuat keseluruhan sistem *text mining* bergantung kepada faktor bahasa. Hal ini menjadi kelemahan dari proses penghilangan *stopword*. Namun proses penghilangan *stopword* tetap digunakan karena proses ini akan sangat mengurangi beban kerja system.

3. Pattern Discovery

Tahap penemuan pola atau *pattern discovery* adalah tahap terpenting dari seluruh proses *text mining*. Tahap ini berusaha menemukan pola atau pengetahuan dari keseluruhan teks. Seperti yang disebutkan dalam bab sebelumnya bahwa dalam data/*text mining* terdapat dua teknik pembelajaran pada tahap *pattern discovery* ini, yaitu *unsupervised* dan *supervised learning*. Adapun perbedaan antara keduanya adalah pada *supervised learning* terdapat label atau nama kelas pada data latih (supervisi) dan data baru diklasifikasikan berdasarkan data latih.

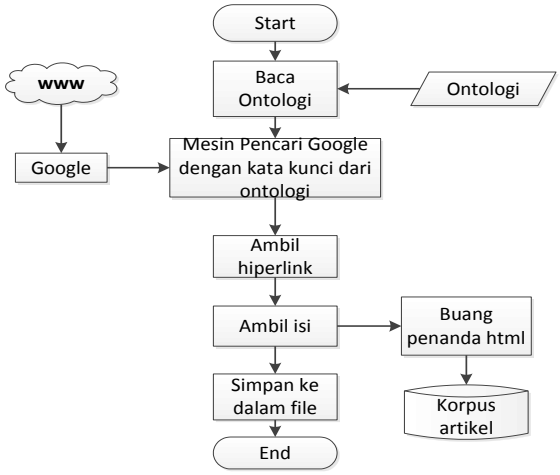
Sedangkan pada *unsupervised learning* tidak terdapat label atau nama kelas pada data latih, data latih dikelompokkan berdasarkan ukuran kemiripan pada suatu kelas. Berdasarkan keluaran dari fungsi, *supervised learning* dibagi menjadi 2, regresi dan klasifikasi. Regresi terjadi

jika output dari fungsi merupakan nilai yang kontinyu, sedangkan klasifikasi terjadi jika keluaran dari fungsi adalah nilai tertentu dari suatu atribut tujuan (tidak kontinyu). Tujuan dari *supervised learning*

adalah untuk memprediksi nilai dari fungsi untuk sebuah data masukan yang sah setelah melihat sejumlah data latih.

ARSITEKTUR SISTEM

Pada gambar 3. dapat dilihat arsitektur sistem Kalsifikasi Berita Menggunakan Ontology yang dibuat dalam penelitian ini.



Gambar 3. Arsitektur Klasifikasi Berita Menggunakan Ontologi

Masing-masing proses dalam arstitektur sistem Klasifikasi Berita menggunakan Ontologi dapat dijelaskan sebagai berikut :

1. Ontologi

Ontologi disimpan ke dalam file Bencana.owl. Untuk membangun ontology digunakan piranti Protégé. Protégé adalah piranti lunak open source yang dikembangkan oleh SMI (Stanford Medical Informatics). Protégé 4.0 yang digunakan dalam penelitian ini adalah piranti untuk mengkonstruksi ontology yang open source, bebas dan memiliki fitur yang terdefinisi dengan baik. Fitur yang paling khusus adalah framework Protégé dibangun sesuai dengan konsep ontology. Protégé mengggunakan multi komponen seperti Protégé-OWL Class, Protégé-Properties, Protégé-Forms, Protégé-Individuals, dan Protégé-Forms, Protégé-Individuals, dan Protégé-OWL Viz untuk mengedit dan membangun ontology, untuk memudahkan perekayasa pengetahuan untuk mengkonstruksi system manajemen pengetahuan berdasarkan ontologi.

Dasar untuk menyusun ontology bencana adalah Undang-Undang no 24 Tahun 2007. Dalam UU no 24 Tahun 2007, jenis bencana ada 3 (tiga) yaitu :

a. Bencana Alam
b. Bencana non alam
c. Bencana social

2. Baca Ontologi

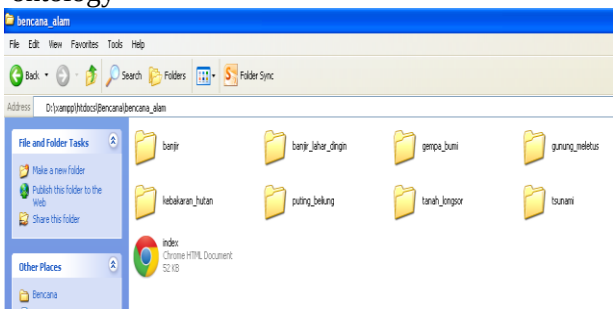
Digunakan untuk membaca file Bencana.owl dan mengubah class dalam ontologi menjadi keyword. Progam ClassHierarchy digunakan untuk membaca file Bencana.owl yang hasilnya disimpan ke dalam variable listm. Listm adalah variabel linked list yang merupakan obyek dari Kelas Listt.

3. Baca Dari Google Search

Hasil pembacaan file Bencana.owl yang disimpan di variable listm digunakan sebagai keyword bagi pencarian Google. Sebagai contoh:

https://www.google.co.id/search?q=bencana+alam&hl=id&btnG=Telusuri

Hasil dari pencarian di google akan disimpan di directory local dengan struktur mengikuti struktur ontology. Pada gambar 4. dapat dilihat directory hasil pencarian dengan menggunakan ontology



Gambar 4 Direktory Local Hasil Pencarian Menggunakan Ontology

Directory ini adah alamat default untuk menampilkan melalui hasil pencarian di google melalui web. Disetiap directory terdapat halaman index.html,index2.html, …index10.html. File-file ini adalah hasil menangkap pencarian google.

File-file ini kemudian dibaca dan diekstrak untuk mendapatkan hyperlink. Dengan memanfaatkan hyperlink ini maka dapat diunduh isi file. Kemudian isi file disimpan di korpus.

PERANCANGAN SYSTEM

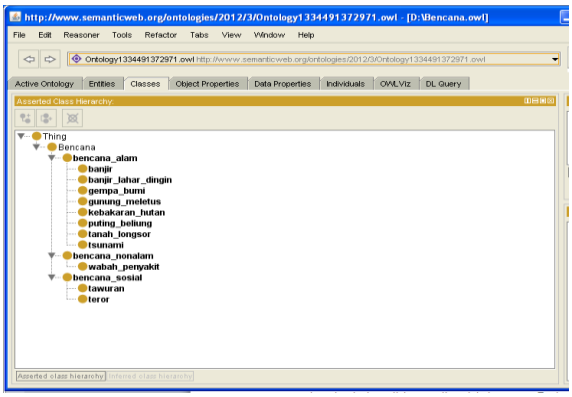
Perancangan sistem terbagi menjadi dua, yaitu perancangan untuk preprosesing dan perancangan ontologi untuk koleksi berita. Perancangan preprosesing menjelaskan tahap persiapan dokumen berita yaitu untuk menyeragamkan bentuk kata, menghilangkan karakter-karakter selain huruf dan mengurangi volume kosa kata. Perancangan ontologi menjelaskan langkah demi langkah pengembangan ontologi mulai dari penentuan domain, istilah/ terminologi, definisi kelas dan hirarki kelas, definisi properti, definisi konstrain dan pembuatan instance.

Perancangan Ontologi

Barikut adalah implementasi untuk membaca file Bencana/owl dan menyimpannya di larik listm. Elemen yang disimpan akan berupa kata dan level dalam pohon ontologi

```
m.read( "file:///D:/Bencana.owl" );
ClassHierarchy Hirar = new ClassHierarchy();
listm = Hirar.showHierarchy( System.out, m );
```

Struktur ontology yang dibuat dengan menggunakan Portege dapat dilihat pada gambar 5.



Gambar 5 Gambar Struktur Ontology Klasifikasi Berita

Selanjutnya dilakukan pembacaan list listm dan diuji jika kedalaman lebih atau sama dengan 2, maka :

1. Jika sama dengan 2, maka buat subdirectory baru di bawah subdirectory Bencana dan simpan file hasil download di subdirectory tersebut..
2. Jika level yang baru sama dengan level saat ini, maka buat subdirectory dengan level yang sama dan simpan file hasil download di subdirectory tersebut.
3. Jika level yang baru sama lebih rendah daripada level saat ini, maka buat subdirectory dengan level yang lebih dalam dan simpan file hasil download di subdirectory tersebut.
4. Jika level yang baru sama lebih tinggi daripada level saat ini, maka buat pindah ke subdirectory dengan level yang lebih tinggi dan simpan file hasil download di subdirectory tersebut.

```
HttpClient httpclient = new DefaultHttpClient();
        String url=word.replaceAll("_","+");
       try {
    HttpGet httpget = new
HttpGet("https://www.google.com/?q=%22"+url+"%22+inurl:te
mpo.co&hl=id&btnG=Telusuri#hl=id&output=search&sclient=
psy-
ab&q=%22"+url+"%22+inurl:tempo.co&oq=&aq=&aqi=&aql=
&gs_l=&pbx=1&bav=on.2,or.r_gc.r_pw.r_cp.r_qf.,cf.osb&fp=7
e8bccf01656ed8f&biw=1280&bih=638");

        System.out.println("executing request " +
httpget.getURI());

        // Create a response handler
        ResponseHandler<String> responseHandler = new
BasicResponseHandler();
        String responseBody = httpclient.execute(httpget,
responseHandler);
        System.out.println("----------------------------------------");

        tulisfile2(direktori+"\\index.htm",responseBody);
```

Prosedur di atas digunakan untuk mendownload halaman www.google.com menggunakan kata kunci yang ada di Bencana.owl. Kata kunci dipassing lewat variabel Word. Kemudian dihilangkan "_" diganti dengan spasi. Fungsi yang digunakan untuk mendownload halaman www.google.com adalah HttpGet. Hasil download akan disimpan di variabel responseBody. Selanjutnya hasil download akan disimpan sesuai directory menggunakan fungsi tulisfile.

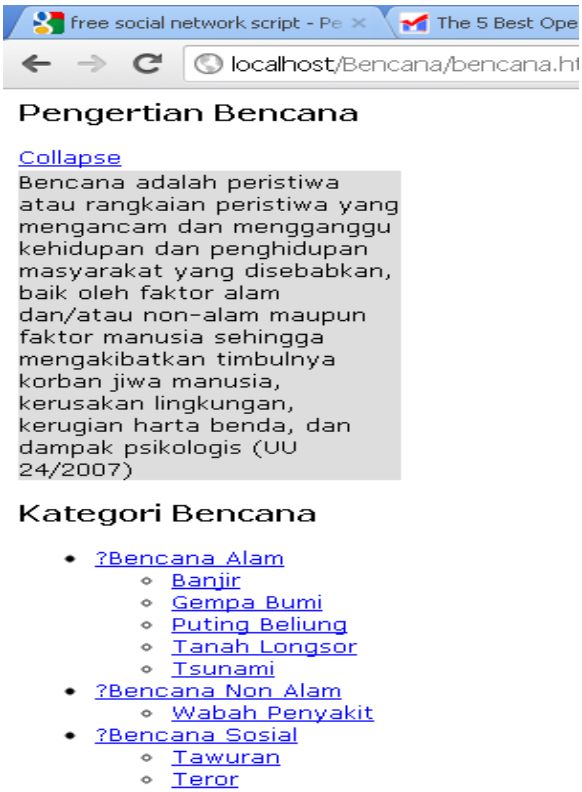
Fungsi tersebut di atas digunakan untuk menyimpan halaman hasil download. Namafile berisi directory dan nama file untuk menyimpan isi hasil download. outFile berisi isi hasil download untuk disimpan ke dalam file.

```
FileWriter outFile = new FileWriter(namafile);
         PrintWriter out = new PrintWriter(outFile);
         // Also could be written as follows on one line
        // Printwriter out = new PrintWriter(new
FileWriter(args[0]));

         // Write text to file
        out.println(isi.toString());
        out.close();
```

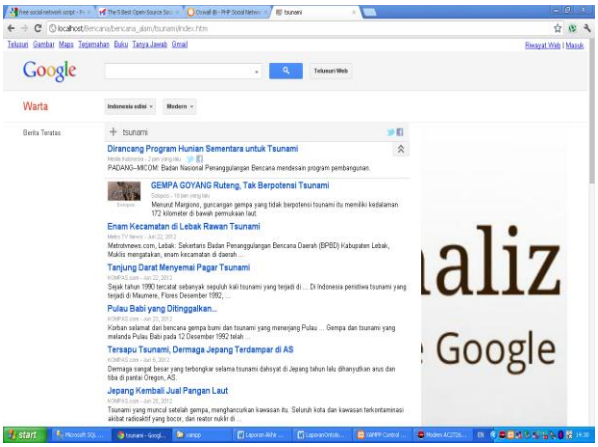
IMPLEMENTASI

Implementasi dari Klasifikasi Berita menggunakan Ontology yang dikembangkan dalam penelitian ini dapat dilihat pada gambar 6 berikut ini :



Gambar 6 Implementasi Klasifikasi Berita Menggunakan Ontology

Jika dilakukan klik subdirektory dari ontology bencana yang telah dibuat, maka proses akan melakukan link dari subdirektory dan menampilkan dokumen hasil link. Tampilan hasil link dapat dilihat pada gambar 7



Gambar 7 Tampilan Hasil Link Subdirektori Ontology

KESIMPULAN

Dalam penelitian ini dapat disimpulkan beberapa hal sebagai berikut :

1. Telah dibuat program **peramban ontologi** untuk mengambil data dari http://news.google.com berdasarkan ontology.
2. Ontology yang disusun berdasarkan Undang-undang no 24 Tahun 2007.
3. Keluaran dari program ini adalah halaman web yang mengandung kata kunci yang tersimpan di file owl.

Dari hasil eksperimen di dapat struktur direktory dan struktur halaman web sesuai dengan struktur ontology

SARAN

Berdasarkan penelitian ini, maka beberapa penelitian yang akan dilakukan berikutnya adalah :

1. Penelitian melakukan klasifikasi berdasarkan ontologi. Dalam penelitian ini akan dilakukan mining agar isi sesuai dengan struktur ontologi yang dibuat.
2. Penelitian melakukan klasifikasi berdasarkan ontologi pada situs mikrobloging twitter. Pada penelitian ini akan dilakukan klasifikasi isi tweet berdasarkan ontology.

DAFTAR PUSTAKA

Baeza-Yates, R., & Ribeiro-Neto, B., 1999, *Modern information retrieval,* NewYork: Addison Wesley.

Coral Calero, dkk., 2006, *Ontologies for Software Engineering and Software Technology*, Springer-Verlag Berlin Heidelberg, New York.

Gruber, T., *What is an Ontology?, http://www-ksl.stanford.edu/kst/what-is-an-ontology.html*

Harlian, Milka. 2006. *Machine Learning Text Kategorization.* Austin : University of Texas.

Hearst, Marti. 2003. *What Is Text Minning?*. SIMS,UC Berkeley. http://www.sims.berkeley.edu/~hearst/text.mining.html . Diakses tanggal 25 Juni 2009.

Horridge., M ., Knublouch, H., et al., 2004, *A Practical Guide to Building OWL Ontologies using the Protégé Owl Plugin co-ode Tool*, edition 1.0 University Manchester & Stanford University.

Khodra, L.M., & Wibisono, Y., 2005, *Clustering berita Berbahasa Indonesia*. Internal Publication, Fakultas Matematika dan Ilmu Pengetahuan Alam, UniversitasPendidikan Indonesia, Bandung, Jawa Barat.

Manning, C., D., Raghavan, P., & Schutze, H., 2008, *Introducion to Information Retrieval*, New York: Cambridge University Press.

Milton, N., 2003, *Knowledge Engineering*, July 21, 2009, http://www.epistemics.so.uk/Notes/61-0-0.htm

Noy., N.F., & McGuinness, D.L, 2001, *Ontology Development 101 : A Guide to Creating Your First Ontology.* Knowledge System Laboratory (KSL) of Departement of Computer Science Stanford USA: Technical Report, KSL-01-05

Pressman R, 2001, *Software Engineering,* Mc Graw Hill, USA.

Salton, M., 1983, *Introduction tomodern information retrieval*, McGraw Hill. New York.

Salton, G., 1989, *Automatic Text Processing, The Transformation, Analysis, and Retrieval of Information by Computer*, Addison – Wesly Publishing Company, Inc. All rights reserved.

Susanto, S., 2006, *Pengklasifikasiandokumen berita menggunakan naïve bayes classifier*, Skripsi, Fakultas Ilmu Komputer,Universitas Indonesia, Depok, Jakarta.

Tenenboim, L., Shapira, B., & Shoval,P., 2008, *Ontology-based classification of news in an electronic news paper Paper presented at Intelligent Information and Engineering Systems Conference,* Varna, Bulgaria

Wibisono, Y., 2005, *Klasifikasi berita berbahasa Indonesia menggunakan naïv ebayes classifier Internal,* Publication,Fakultas Matematika dan Ilmu PengetahuanAlam, Universitas Pendidikan Indonesia,Bandung, Jawa Barat.

_____,ONTOLOGI: Bahasa dan Tools PROTÉGÉ, *http://paperwgdbis.abmutiara.info/tutorial/Bahasa_tool_ontology.pdf*, 5.2009