BAB I

PENDAHULUAN

1.1. LATAR BELAKANG

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. Web memuat banyak informasi yang dihasilkan dari waktu ke waktu secara kontinue dari berbagai sumber. Jumlah informasi yang terus bertambah dari waktu ke waktu dapat menyulitkan para pencari informasi dalam menemukan informasi yang relevan. 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. Dengan meningkatnya kebutuhan para pembaca berita untuk mendapatkan berita-berita yang terkait dengan berita yang dibacanya saat ini, baik dari sisi topik ataupun kejadian dalam berita tersebut. 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.

Berdasarkan latar belakang di atas, maka dalam penelitian ini akan dibangun aplikasi Klasifikasi Berita Menggunakan Ontologi. Diharapkan aplikasi yang dibuat akan lebih menghemat waktu dan memudahkan pencarian informasi yang sesuai (relevan) dengan yang diinginkan oleh pengguna.

1.2. PERUMUSAN MASALAH

Berdasarkan latar belakang di atas, maka permasalahan yang dapat dirumuskan adalah :

1. Bagaimana mengekstrak dokumen teks dari suatu koleksi dokumen berita dari halaman web
2. Bagaimana melakukan klasifikasi dokumen berita menggunakan ontologi.

1.3. BATASAN MASALAH

Dalam penelitian ini ada beberapa pembatasan masalah yang dilakukan, yaitu:

1. Data yang digunakan dalam penelitian ini berupa artikel berita berbahasa Indonesia dari situs http://new.google.com.
2. Dalam penelitian ini menggunakan domain bencana alam untuk melakukan klasifikasi dokumen.
3. Klasifikasi dokumen dengan menggunakan ontologi dilakukan dengan membandingkan nilai kemiripan diantara dokumen dan sebuah node yang ada di ontologi.

BAB II

TUJUAN DAN MANFAAT PENELITIAN

2.1. TUJUAN PENELITIAN

Tujuan yang ingin dicapai dalam penelitian ini adalah :

1. Merancang dan membuat aplikasi untuk membaca dokumen berita dari hasil pencarian www.google.co.id.
2. Merancang dan membuat aplikasi yang dapat membuat struktur ontologi untuk klasifikasi dokumen berita.

2.2. MANFAAT PENELITIAN

Manfaat yang diharapkan dari penelitian ini adalah :

1. Hasil penelitian bisa digunakan untuk melakukan penyimpanan dokumen berita dari web yang telah diklasifikasikan berdasarkan jenis.
2. Penelitian ini dapat digunakan untuk pencarian dokumen relevan berdasarkan kata kunci tertentu yang diinputkan pengguna dan mempersingkat waktu pencairan berita dengan topik tertentu.
3. Dapat menjadi rujukan bagi penelitian selanjutnya yang memiliki keterkaitan dengan penelitian secara langsung maupun tidak langsung

BAB III

TELAAH PUSTAKA

3.1. 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". Christopher D. Manning dan rekannya (Manning, C., D., Raghavan, P., 2008) memberikan definisi tentang perolehan informasi, yaitu "pencarian informasi, biasanya berupa dokumen, dari sesuatu yang tidak terstruktur, biasanya berupa teks, dalam suatu koleksi yang dapat memenuhi kebutuhan informasi yang diinginkan".

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. Untuk menanggulangi masalah ini dibutuhkan suatu sistem yang dapat memudahkan pencari informasi untuk mendapatkan informasi yang diinginkan dari suatu koleksi dokumen. Sistem ini kemudian dikenal dengan nama sistem perolehan informasi. Sistem perolehan informasi merupakan sistem yang diharapkan dapat memberikan sebanyak-banyaknya informasi dan serelevan mungkin terhadap kebutuhan pengguna sistem tersebut. Sebuah sistem perolehan informasi memiliki beberapa operasi dasar, seperti pembuatan indeks koleksi dokumen dan pencarian dokumen berdasarkan kueri yang dimasukkan oleh pencari informasi. Kueri tersebut kemudian dicocokkan dengan koleksi dokumen yang ada di dalam *database*. Dokumen yang ditampilkan berdasarkan queri masukan belum terurut dari dokumen yang paling relevan sampai yang tidak relevan sehingga dokumen tersebut perlu dilakukan klasifikasi. Klasifikasi dokumen merupakan topik utama yang akan dijelaskan pada penelitian ini.

3.2. 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. Hasil eksperimen dalam penelitian ini menunjukkan bahwa jumlah dokumen pembelajaran 90% dan jumlah dokumen

pembelajaran 90% dan jumlah dokumen pengujian 10% (stemming) menghasilkan akurasi yang paling tinggi yaitu dengan recall 93,5%, precision 90,36%, dan f-measure 93,81%. Kesimpulan yang diperoleh dari penelitian ini adalah kinerja Naïve Bayes classifier yang menggunakan stemming lebih baik dari pada non-stemming. Penelitian dalam bidang klasifikasi dokumen berbahasa Indonesia yang dilakukan oleh Yudi Wibisono dan Slyvia Susanto membutuhkan dokumen pembelajaran untuk melakukan klasifikasi dokumen baru. Pada paper ini mengusulkan sebuah metode baru untuk melakukan klasifikasi dokumen, yaitu dengan menggunakan ontologi. Metode klasifikasi dokumen dengan menggunakan ontologi tidak memerlukan dokumen pembelajaran

3.3. 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

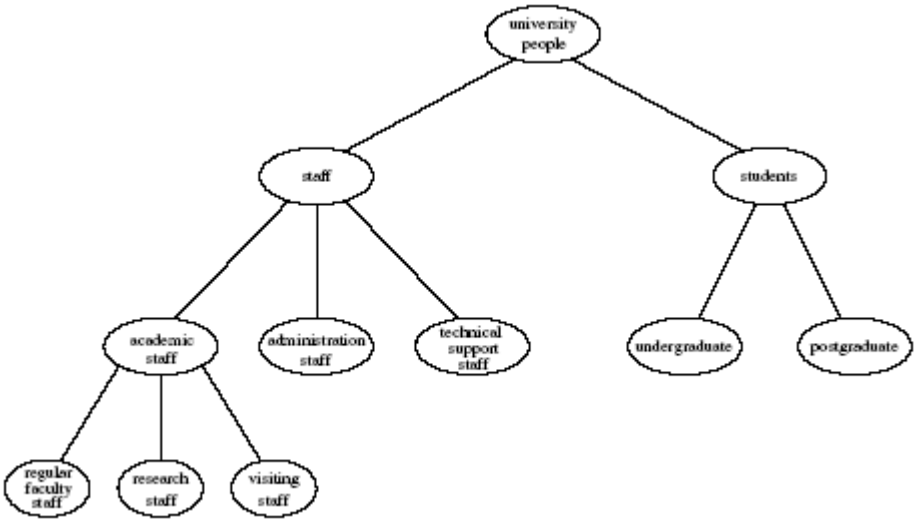
memahami domain yang digunakan untuk klasifikasi dokumen dengan baik. Akan tetapi, hal inilah yang menjadi kelemahan *rule base*, yaitu metode klasifikasi dokumen menggunakan ontologi sangat bergantung pada adanya pakar. Selain itu, pendekatan ini memiliki kekurangan lain, yaitu membutuhkan waktu yang 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.

3.4. 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.

Menurut definisi T.R. Gruber's, yang kemudian diperbaiki R. Studer: Sebuah ontologi adalah spesifikasi yang eksplisit dan formal dari sebuah konseptualisasi. 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 3.1. Contoh Ontologi Domain Universitas

Disamping menujukkan hubungan subkelas, ontologi juga dapat mencakup informasi seperti :

- Property (X mengajar Y)
- Batasan harga (hanya staff akademik yang dapat mengajar matakuliah)
- Pernyataan disjoin (staff akademik dan administrasi disjoin)
- Spesifikasi hubungan logika diantara objek (setiap departemen harus mencakup paling tidak sepuluh anggota staff akademik)

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.

Pemakai mungkin mengklik salah satu pilihan untuk memperluas subkategori. Ontologi juga meningkatkan keakuratan pencarian Web. Mesin pencari dapat mencari halaman-halaman yang menunjuk ke konsep yang tepat dalam sebuah ontologi. Mesin pencari web juga dapat mengeksploitasi informasi generalisasi atau spesialisasi. Jika suatu *query* gagal untuk menemukan dokumen yang relevan,mesin pencari dapat menyarankan pemakai sebuah *query* yang lebih

general. Atau untuk mencegah terlalu banyak jawaban yang diberikan, maka mesin pencari dapat menyarankan pencarian yang khusus (spesialisasi).

3.4.1. 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*.

Ontologi baru dapat digunakan apabila ontologi tersebut sudah diekspresikan terlebih dahulu dalam notasi yang nyata. Sebuah bahasa ontologi adalah sebuah bahasa formal yang digunakan untuk merepresentasikan ontologi. Beberapa komponen bahasa yang menyusun ontologi, yaitu XML, XML *schema*, RDF, RDF *schema*, dan *Ontology Web Language* (OWL). XML menyediakan sintaksis untuk dokumen keluaran secara terstruktur, tetapi belum menggunakan *semantic constraints*. XML *schema* adalah bahasa untuk pembatasan struktur dari dokumen XML. RDF adalah model data untuk objek dan relasi diantaranya, menyediakan semantic yang sederhana untuk model data, dan disajikan dalam sintaks XML. RDF *schema* adalah kosakata untuk menjelaskan properti dan *class* dari sumber RDF. OWL adalah bahasa ontologi yang baru untuk sebuah web semantik, dikembangkan oleh *World Wide Web Consortium* (W3C) Horridge., M ., Knublouch, H., 2004) OWL dapat mendefinisikan relasi antar *class*, kardinalitas, karakteristik dari *properties*, dan *equality*. Ontologi dapat digunakan untuk

melakukan klasifikasi dokumen teks dalam penelitian ini karena ontologi bersifat unik dan memiliki struktur hierarkis. Selain itu, sebuah model ontologi dapat menghilangkan makna ambigu, sehingga dapat menanggulangi masalah yang muncul pada bahasa alami di mana sebuah kata memiliki lebih dari satu makna atau arti bergantung pada konteks kalimatnya. Pengembangan ontologi dalam penelitian ini terdiri dari beberapa komponen utama yaitu:

1. Konsep

   Konsep atau *class* merepresentasikan *term* atau kata dalam domain yang spesifik.

2. Fitur

   Fitur atau *instance* merepresentasikan individu dari sebuah kelas.

3. Relasi

   Relasi atau *property* merepresentasikan hubungan diantara konsep. Ada dua relasi yang digunakan dalam penelitian ini yaitu: relasi "*is-a*" dan "*has-a*".

4. *Constraint*

   Constraint merepresentasikan kondisi yang harus dipenuhi di sebuah konsep.

3.4.2. 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).

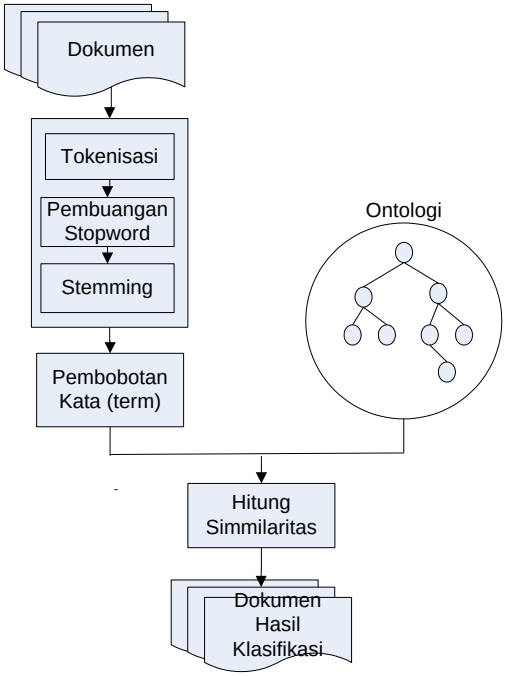
Tujuan dari proses persiapan dokumen teks adalah untuk menghilangkan karakter-karakter selain huruf, menyeragamkan kata dan mengurangi volume kosa kata. Proses pembobotan kata adalah proses memberikan nilai atau bobot ke sebuah kata berdasarkan kemunculannya pada suatu dokumen teks (Baeza-Yates, R.,1999). 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).

Term frequency – inverse document frequency atau biasa sering disebut TF-IDF adalah metode pembobotan kata dengan menghitung nilai TF dan juga menghitung kemunculan sebuah kata pada koleksi dokumen teks secara keseluruhan (Baeza-Yates, R., 1999). Pada pembobotan ini, jika kemunculan term pada sebuah dokumen teks tinggi dan kemunculan term tersebut pada dokumen teks lain rendah, maka bobotnya akan semakin besar. Akan tetapi, jika kemunculan term tersebut pada dokumen teks lain tinggi, maka bobotnya akan semakin kecil. 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).

$$tfidf(i,j) = tf(i,j) \times \log\left(\frac{N}{df(j)}\right) \quad (1)$$

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. Proses klasifikasi dokumen dengan menggunakan ontologi dilakukan setelah melakukan pembobotan kata. Proses klasifikasi dilakukan dengan memetakan dokumen teks ke sebuah node dengan nilai kemiripan paling tinggi dan dokumen teks tersebut diklasifikasikan tepat ke satu class. Rumus untuk menghitung nilai kemiripan dapat dilihat pada Persamaan (2 dimana N adalah frekuensi fitur dari sebuah node, freq i,d merepresentasikan frekuensi fitur dari fitur i yang cocok di dokumen d, max i,d merepresentasikan frekuensi fitur yang paling cocok di dokumen d, V adalah jumlah constraint, dan V d adalah jumlah constraint yang terpenuhi di dokumen d. Proses klasifikasi dokumen hanya dilakukan ketika menggunakan relasi “ is-a ” dan “ has-a”. Ketika node lain cocok dengan fitur di dokumen, maka node tersebut juga dimasukkan ke dalam proses klasifikasi dokumen untuk menghitung nilai kemiripannya. Perancangan klasifikasi dokumen teks dengan menggunakan ontologi dapat dilihat pada Gambar 3.2



Gambar 3.2. Perancangan Klasifikasi Dokumen Menggunakan Ontologi

3.3. 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.

3.3.1. Tahapan *Text Mining*

Walaupun inti dari suatu sistem klasifikasi adalah tahap penemuan pola (*pattern discovery*) namun secara lengkap proses *text mining* dibagi menjadi 3 tahap utama, yaitu proses awal terhadap teks (*text preprocessing*), transformasi teks ke dalam bentuk antara (*text transformation/feature generation*), dan penemuan pola (*pattern discovery*). Masukan awal dari proses ini adalah suatu data teks dan menghasilkan keluaran berupa pola sebagai hasil interpretasi.

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.

Proses *partofspeech* melakukan parsing terhadap seluruh kalimat dalam teks kemudian memberikan peran kepada setiap kata, misalnya : petani (subyek)

pergi (predikat) ke (kata hub) sawah (keterangan). Hasil dari *partofspeech tagging* dapat digunakan untuk *parse tree*, di mana masing-masing kalimat berdiri sebagai sebuah pohon mandiri. Untuk proses parsing sederhana tidak dibangun *parse tree* seperti cara sebelumnya. Pada proses parsing sederhana sistem akan memecah teks menjadi sekumpulan kata-kata, yang kemudian akan dibawa sebagai input untuk tahap berikutnya pada proses *text mining*.

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. Dengan menghilangkan *stopword* dari suatu teks maka sistem hanya akan memperhitungkan kata-kata yang dianggap penting. *Stemming* adalah contoh tindakan lain yang dapat dilakukan pada tahap transformasi teks. *Stemming* adalah proses untuk mereduksi kata ke bentuk dasarnya Sedangkan menurut Tala (2003) *Stemming* adalah suatu proses yang menyediakan suatu pemetaan antara berbagai kata dengan morfologi yang berbeda menjadi satu bentuk dasar (*stem*). Kata yang memiliki bentuk dasar sama

walaupun imbuhannya berbeda seharusnya memiliki kedekatan arti. Disamping itu juga, proses *stemming* akan sangat mengurangi jumlah dan beban database. Jika setiap kata disimpan tanpa melalui proses stemming, maka satu macam kata dasar saja akan disimpan dengan berbagai macam bentuk yang berbeda sesuai dengan imbuhan yang mungkin melekatinya. Hal ini sangat berbeda jika kita menerapkan proses *stemming* pada tahap ini, satu kata dasar hanya akan disimpan sekali walaupun mungkin kata dasar tersebut pada sumber data sudah berubah dari bentuk aslinya dan mendapatkan berbagai macam imbuhan. Proses *stemming* dan penghilangan *stopword* dapat digunakan secara mandiri atau tergabung, dimana dilakukan proses penghilangan *stopword* terlebih dahulu yang diikuti dengan proses *stemming*. Hal ini dilakukan untuk menemukan pola dari teks dalam berita tersebut.

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.

Berikut tahapan umum yang biasa dilakukan pada *supervised learning*:

1. Menentukan tipe data latih.
2. Mengumpulkan data latih. Data latih yang digunakan seharusnya memiliki karakteristik dunia nyata. Karena itu data latih dapat berasal baik dari hasil pengukuran atau dari pakar.
3. Menentukan representasi fitur masukan dari fungsi yang ingin dibentuk karena tingkat akurasi dari fungsi dapat dipengaruhi oleh representasi dari masukan.
4. Menentukan struktur dari pengetahuan (fungsi) dan algoritma yang akan digunakan.
5. Jalankan algoritma terhadap data latih.

BAB IV

METODE PENELITIAN

4.1. OBYEK 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.

4.2. TEKNIK PENGUMPULAN DATA

Pengumpulan data dimaksudkan agar mendapatkan bahan-bahan yang relevan, akurat dan reliable. Maka teknik pengumpulan data yang dilakukan dalam penelitian ini adalah sebagai berikut :

1. **Observasi**

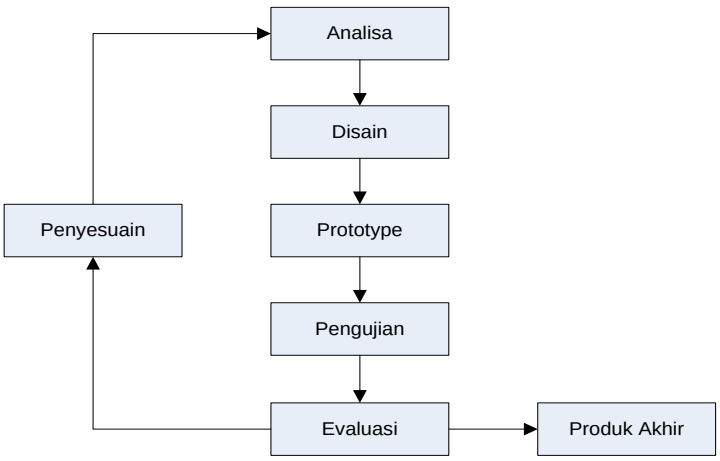
   Dengan melakukan pengamatan dan pencatatan secara sistematis tentang hal-hal yang berhubungan dengan basis data dokumen teks dan kemampuan pencarian kemiripan dokumen.

2. **Studi Pustaka**

   Dengan pengumpulan data dari bahan-bahan referensi, arsip, dan dokumen yang berhubungan dengan permasalan dalam penelitian ini.

4.3. METODE PENGEMBANGAN

Penelitian ini menggunakan model *prototyping.* Di dalam model ini sistem dirancang dan dibangun secara bertahap dan untuk setiap tahap pengembangan dilakukan percobaan-percobaan untuk melihat apakah sistem sudah bekerja sesuai dengan yang diinginkan. Sistematika model *prototyping* terdapat pada Gambar 4.1 memperlihatkan tahapan pada prototyping.



Gambar 4.1 Tahapan Prototyping (Pressman, 2001)

Berikut adalah tahapan yang dilakukan pada penelitian ini dengan metode pengembangan prototyping

1. **Analisa**

   Pada tahap ini dilakukan analisa tentang masalah penelitian dan menentukan pemecahan masalah yang tepat untuk menyelesaikannya.

2. **Disain**

   Pada tahap ini dibangun rancangan sistem dengan beberapa diagram bantu seperti Data Flow Diagram.

3. **Prototype**

   Pada tahap ini dibangun aplikasi berbasis web yang sesuai dengan disain dan kebutuhan sistem.

4. **Pengujian**

   Pada tahap ini dilakukan pengujian pada pustaka fungsi yang sudah dibangun.

5. **Evaluasi**

   Pada tahap ini dilakukan evaluasi apakah performa aplikasi sudah sesuai dengan yang diharapkan, apabila belum maka dilakukan penyesuaian-penyesuaian secukupnya.

6. **Penyesuaian**

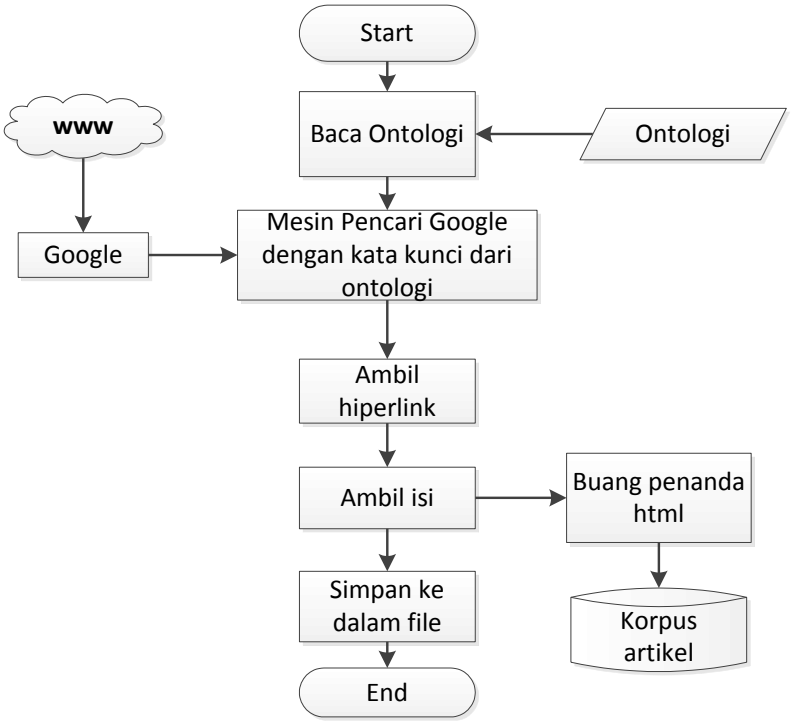
   Tahap ini dilakukan apabila pada evaluasi performa aplikasi kurang memadai dan dibutuhkan perbaikan, tahap ini melakukan penyesuaian dan perbaikan pada aplikasi sesuia dengan kebutuhan

BAB V

HASIL DAN PEMBAHASAN

5.1. ARSITEKTUR SISTEM

Arsitektur sistem merupakan diagram yang menggambarkan urutan proses-proses dari sistem yang dibangun. Pada gambar 5.1 dapat dilihat arsitektur sistem Kalsifikasi Berita Menggunakan Ontology yang dibuat dalam penelitian ini.



Gambar 5.1. 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

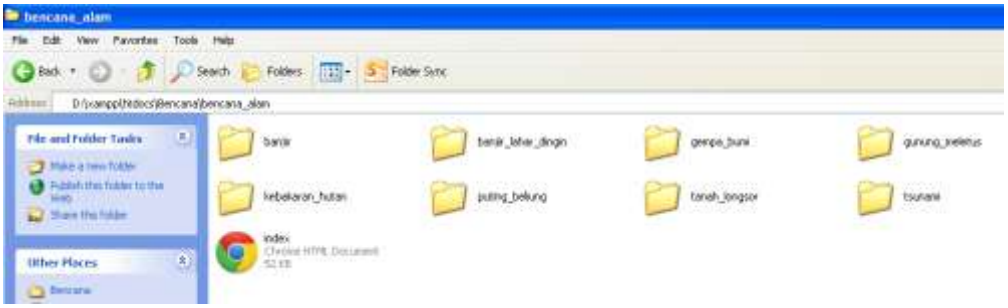
file Bencana.owl yang hasilnya disimpan ke dalam variable listm. Listm adalah variabel linked list yang merupakan obyek dari Kelas Listt.

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 5. dapat dilihat directory hasil pencarian dengan menggunakan ontology



Gambar 5.2 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.

5.2. 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.

5.2.1. PERANCANGAN ONTOLOGI

Noy dan McGuinness (2000) telah menjelaskan ada beberapa langkah-langkah yang harus diperhatikan didalam pengembangan ontologi, salah satunya dengan menentukan konsep dan domain.

Penentuan Konsep dan Domain

Penentuan konsep dan domain pada dasarnya merepresentasikan koleksi semua dokumen yang dilengkapi dengan informasi dan disusun berdasar suatu klasifikasi dan dikelompokkan kedalam jenis-jenis yang sama (Class). Dokumen tersebut meliputi article, proceedings, textbook, periodicreport (journal, magazine, newspaper) dan finalproject (bachelorthesis, mastersthesis, phdthesis).

Penentuan Daftar Terminologi

Penentuan daftar terminologi menegaskan hal-hal yang berkaitan dengan istilah-istilah yang digunakan didalam membuat statemen sekaligus memberikan jawaban dari statemen yang dibuat sebelumnya.

1. Definisi Kelas dan Struktur Hirarki

Representasi definisi kelas dan hirarki kelas adalah mengelompokkan kelas-kelas dengan karakteristik yang sama yang muncul didalam sebuah domain. Uschold dan Gruninger (1996) menegaskan didalam makalah " Ontologies: Principles, Methods and Aplicatons" bahwa ada beberapa pendekatan metode yang dapat digunakan untuk membangun struktur hirarki kelas diantaranya yaitu metode top-down, bottom-up, combination. Untuk aplikasi bibliografi ini digunakan pendekatan top-down, dimana kelas-kelas didefinisikan dari mulai konsep yang paling umum sampai konsep yang lebih spesifik.

2. Diagram Kelas

Diagram kelas atau class diagram menunjukkan interaksi(relasi) antar kelas didalam sistem, sehingga dengan diagram kelas penyajian informasi yang dimiliki oleh setiap kelas dapat terlihat jelas.

3. Definisi Properti (Slot)

Setelah kelas diciptakan langkah selanjutnya adalah mendefinisikan properti kelas. Sebuah kelas jika berdiri sendiri tidak akan memberikan informasi yang cukup tanpa adanya properti yang melekat didalamnya. Dengan properti inilah sebuah kelas akan mempunyai nilai tambah dalam hal ini adalah informasi.

4. Konstrain Properti

Konstrain properti merupakan batasan tertentu dimana properti yang dimiliki setiap kelas memiliki tipe nilai khusus. Didalam pengembangan ontologi konstrain properti dikategorikan menjadi dua kategori.

a. Slot Kardinalitas, Slot kardinalitas didefinisikan sebagai nilai banyaknya yang dimiliki setiap properti kelas.
b. Slot Tipe, Slot tipe menegaskanbeberapa tipe data properti yang harus didefinisikan.

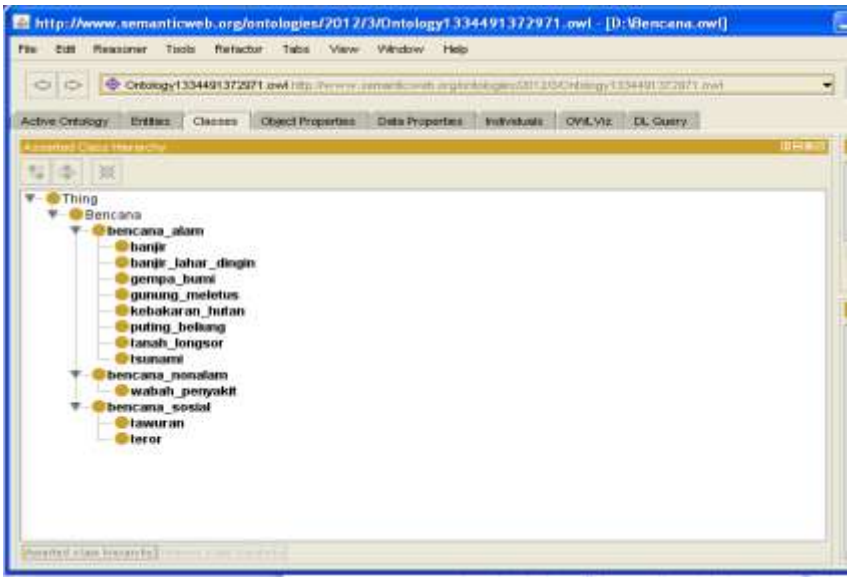
5. Pembuatan Instance

Langkah terakhir setelah konsep pembuatan properti kelas selanjutnya menciptakan sebuah instance dari kelas. Langkah pendefinisian sebuah instance kelas dimulai dengan memilih kelas, membuat individu instance dari kelas kemudian yang terakhir mengisi slot properti dari individu instance kelas.

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 5.3.



Gambar 5.3. Gambar Struktur Ontology Klasifikasi Berita

Dari struktur ontologi yang telah dibuat menggunakan Portege dengan struktur sebagai berikut :

0 Bencana
- 1 bencana_alam
  - 2 banjir
  - 2 banjir_lahar_dingin
  - 2 gempa_bumi
  - 2 gunung_meletus
  - 2 kebakaran_hutan
  - 2 puting_beliung
  - 2 tanah_longsor
  - 2 tsunami
- 1 bencana_non_alam
  - 2 wabah_penyakit
- 1 bencana_sosial
  - 2 tawuran
  - 2 teror

Selanjutnya adalah implementasi untuk pembacaan list listm dan pengujian kata kunci yang diambil dari struktur ontology yang telah dibuat adalah sebagai berikut :

```
while (listm.isEmpty() != true)
{
                owlstring.append(listm.removeFromFront());
                listdepth.add(owlstring.substring(0,1).toString());
               owlstring.deleteCharAt(0);
               listowl.insertAtBack(owlstring.deleteCharAt(0).toString());

              //penelusuran
               int a=0;
               int b=0;
               int c=1;
               if (listdepth.size()>=2)
```

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

```
HttpClient httpclient = new DefaultHttpClient();
                        String url=word.replaceAll("_","+");
      try {
            HttpGet httpget = new
HttpGet("https://www.google.com/?q=%22"+url+"%22+inurl:tempo.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=7e8bccf01656ed
8f&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);
```

buat pindah ke subdirectory dengan level yang lebih tinggi dan simpan file hasil download di subdirectory tersebut.

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.

```
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();
```

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.

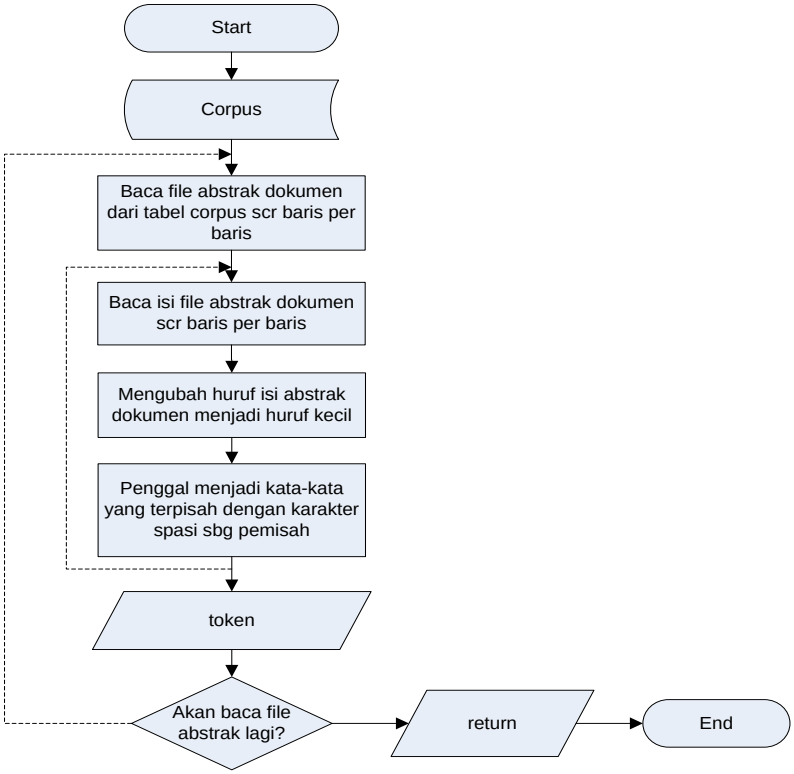
5.2.2. PERANCANGAN PREPROSESING

Perancangan preprosesing menjelaskan proses ataupun langkah-langkah yang dilakukan dalam persiapan dokumen berita yang akan digunakan sebagai data untuk klasifikasi. Persiapan dokumen ini dimaksudkan untuk menyeragamkan bentuk kata, menghilangkan karakter-karakter selain huruf dan

mengurangi volume kosa kata. Langkah-langkah dalam persiapan dokumen berita ini adalah sebagai berikut : tokenisasi *(tokenizations),* pembuangan stopword *(stopword removal*), pembentukan kata dasar *(stemming)*

Tokenisasi

Sebelum kata dipisahkan dari kalimatnya, terlebih dahulu dibersihkan dari tanda baca, tag html dan angka. Pada penelitian ini untuk membersihkan tanda baca dapat digunakan perintah yang disediakan oleh Java. Pembersihan dilakukan sebelum proses tokenisasi (*tokenizations)* dimaksudkan untuk memperkecil hasil dari tokenisasi. Pada proses tokenisasi akan dibaca dokumen abstrak dalam format teks akan dilakukan proses pemotongan string input berdasarkan tiap kata yang menyusunnya. Pada umumnya setiap kata terindentifikasi atau terpisahkan dengan kata yang lain oleh karakter spasi, sehingga proses tokenisasi mengandalkan karakter spasi pada dokumen untuk melakukan pemisahan kata.



Gambar 5. 4 Flowchart Proses Tokenisasi

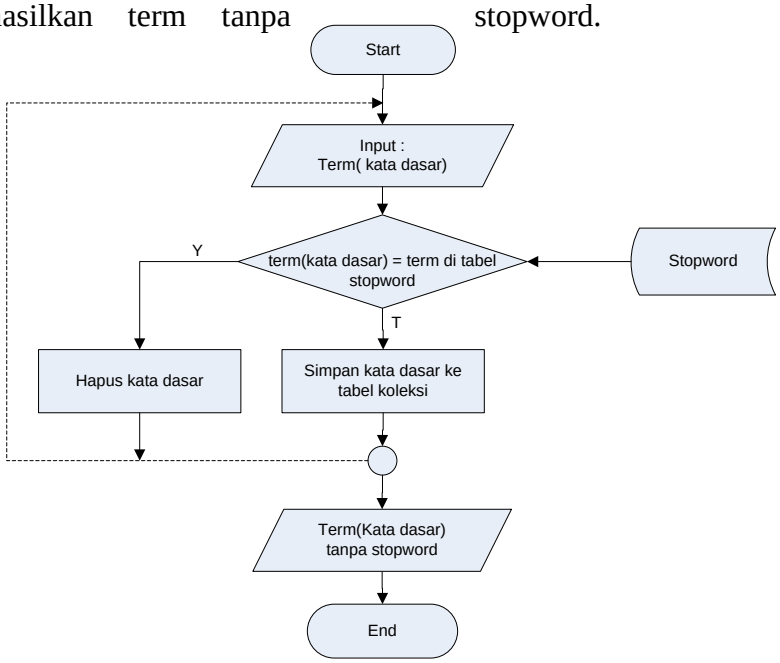
Seperti yang terlihat pada gambar 5.4 pada proses preprosesing untuk tokenisasi, semua term dalam dokumen yang dibaca diganti dengan huruf kecil. Setelah itu tiap term akan dicek apakah tanda baca atau tidak. Jika tanda baca maka akan dihapus/dibuang. Proses akan dilanjutkan untuk membuat *term* menjadi token-token yang terpisah.

Pembuangan Stopword

Proses pembuangan stopword dimaksudkan untuk mengetahui suatu kata masuk ke dalam stopword atau tidak. Pembuangan stopword adalah proses pembuangan term yang tidak memiliki arti atau tidak relevan. Term yang diperoleh dari tahap tokenisasi dicek dalam suatu daftar *stopword*, apabila sebuah kata masuk di dalam daftar stopword maka kata tersebut tidak akan diproses lebih lanjut. Sebaliknya apabila sebuah kata tidak termasuk di dalam daftar stopword maka kata tersebut akan masuk keproses berikutnya. Daftar stopword tersimpan dalam suatu tabel, dalam penelitian ini menggunakan daftar stop word yang digunakan oleh Tala (2003), yang merupakan stopword Bahasa Indonesia yang berisi kata-kata seperti ; ini, itu, yang, ke, di, dalam, kepada, dan seterusnya sebanyak 780 kata.

Seperti terlihat pada gambar 5.5 pembuangan stopword dilakukan dengan mengecek pada tabel stopword. Bila term cocok dengan salah satu isi tabel

stopword, maka term tersebut dianggap sebagai stopword akan dibuang dan tidak akan diikutkan pada proses *stemming*. Dari proses pembuangan stopword akan menghasilkan term tanpa stopword.



Gambar 5. 5 Flowchart Proses Pembuangan Stopword

Stemming

Proses *stemming* adalah proses pembentukan kata dasar. Term yang diperoleh dari tahap pembuangan stopword akan dilakukan proses stemming. Algoritma stemming yang digunakan adalah modifikasi Porter stemmer dari (Tala, 2003). Stemming digunakan untuk mereduksi bentuk term untuk menghindari ketidak cocokan yang dapat mengurangi recall, di mana term-term yang berbeda namun memiliki makna dasar yang sama direduksi menjadi satu bentuk.

Proses stemming adalah bagian dari proses preprosesing, tujuan utama dari proses stemming adalah mengembalikan kata dalam bentuk dasarnya. Dengan kata dasar dapat mereduksi bentuk term untuk menghindari ketidak cocokan yang

dapat mengurangi recall, di mana term-term yang berbeda namun memiliki makna dasar yang sama direduksi menjadi satu bentuk.

Struktur pembentukan kata dalam Bahasa Indonesia adalah sebagai berikut:

[awalan-1] + [awalan-2] + dasar + [akhiran] + [kepunyaan] + [sandang]

Masing-masing bagian tersebut (yang dalam kotak bisa ada atau tidak), digabungkan dengan kata dasar membentuk kata berimbuhan.

Penggunaan algoritma stemming Tala bertujuan untuk mempercepat waktu implementasi dan diharapkan performa yang stabil walaupun data dokumen bertambah terus. Algoritma Tala menggunakan algoritma rule based stemming seperti halnya dengan algoritma porter pada stemming bahasa Inggris.

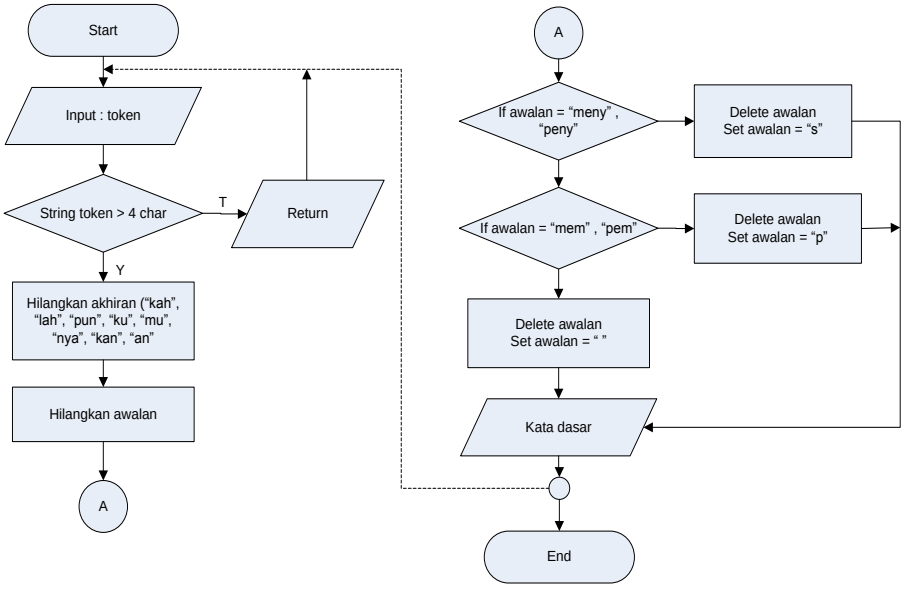
Pada stemmer Tala terdapat 5 langkah utama dengan 3 langkah awal dan 2 langkah pilihan, langkah-langkah tersebut sbb:

a) Menghilangkan partikel
b) Menghilangkan kata sandang dan kepunyaan.
c) Menghilangkan awalan 1
d) Jika suatu aturan terpenuhi jalankan sbb :
    - Hilangkan Akhiran
    - Jika suatu aturan terpenuhi, hilangkan awalan 2. Jika tidak proses stemming selesai
e) Jika tidak ada aturan yang terpenuhi jalankan sbb :
    - Hilangkan awalan 2.
    - Hilangkan Akhiran

- Proses stemming selesai.

Selain itu tala juga membagi imbuhan menjadi 5 cluster yang nantinya digunakan untuk menghilangkan imbuhan pada setiap tahapnya.

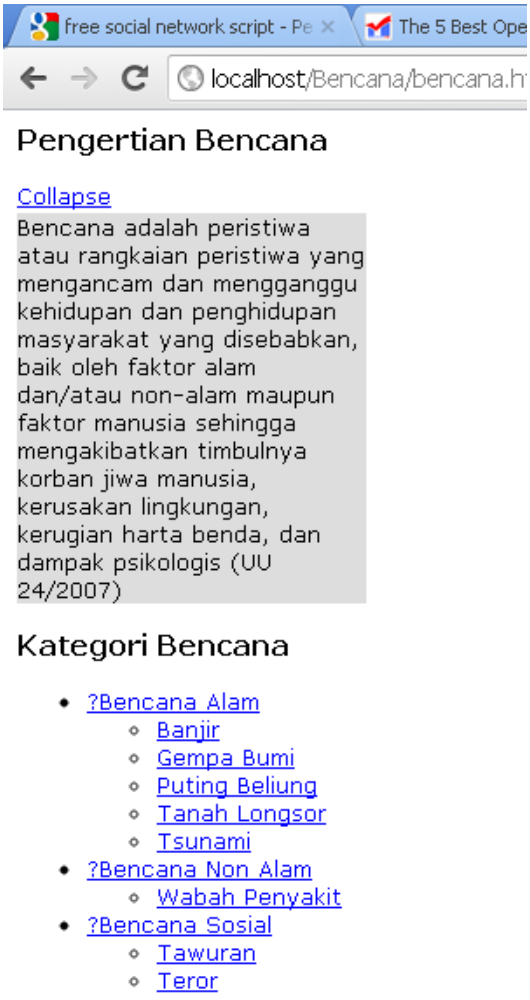
Dapat dilihat pada gambar 5.6 tahap pertama proses stemming adalah mengecek jumlah karakter lebih besar dari 4, jika karakter lebih dari 4 akan dilakukan proses menghilangkan kata sandang dan kepunyaan. Setelah proses berhasil dilakukan akan dilanjutkan proses menghilangkan awalan. Sebelum proses menghilangkan awalan dilakukan akan dicek terlebih dahulu apakah karakter lebih dari 3. Jika tidak maka akan disimpan sebagai kata dasar. Jika karakter lebih dari 3 akan dicek apakah awalan adalah string "meny", "peny" jika benar maka dihasilkan kata dasar dengan karakter awal diganti dengan karakter "s". Jika tidak akan dicek apakah awalan adalah string "mem", "pem". Jika benar maka akan dihasilkan kata dasar dengan karakter awal diganti dengan karakter "p". Jika awalan tidak string tersebut (meny, peny, mem, pem) maka awalan akan dihilangkan dan akhir proses akan dihasilkan kata dasar.



Gambar 5.6 Flowchart Proses Stemming

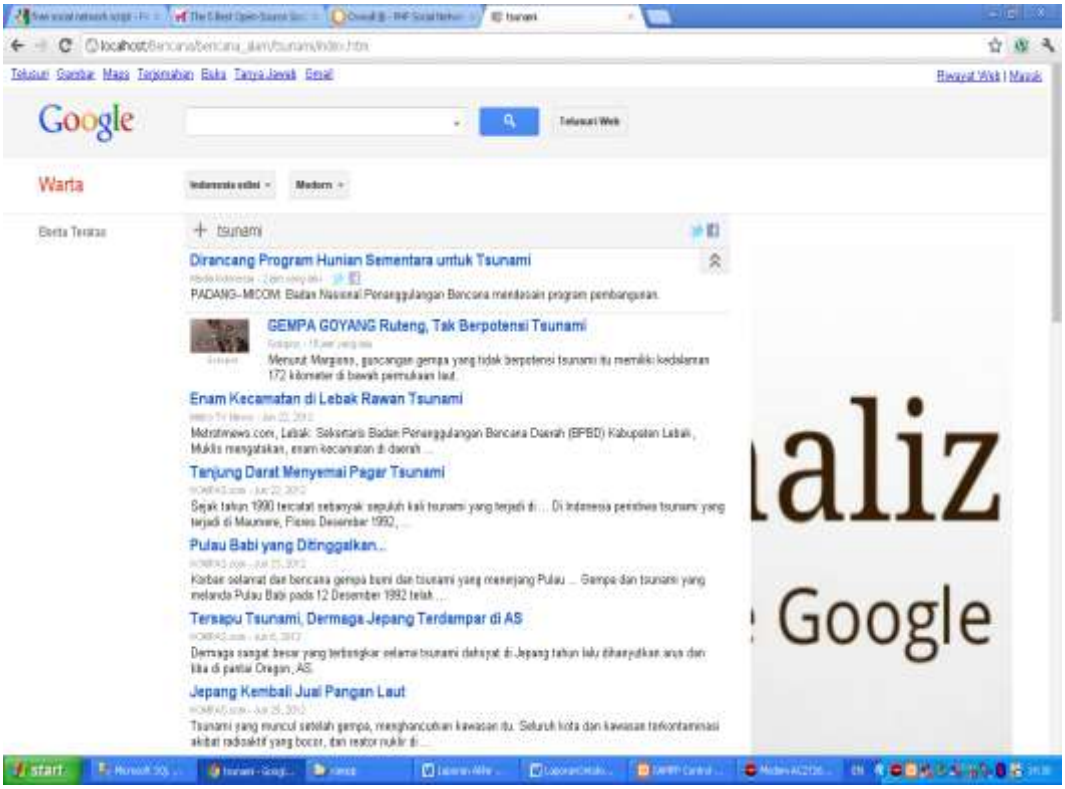
5.3. IMPLEMENTASI

Implementasi dari Klasifikasi Berita menggunakan Ontology yang dikembangkan dalam penelitian ini dapat dilihat pada gambar 5.7 berikut ini :



Gambar 5.7 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 5.8.



Gambar 5.8 Tampilan Hasil Link Subdirektori Ontology

BAB VI

KESIMPULAN DAN SARAN

6.1. 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.
4. Dari hasil eksperimen di dapat struktur direktory dan struktur halaman web sesuai dengan struktur ontology.

6.2. Penelitian Berikutnya

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 ontologi.

DAFTAR PUSTAKA

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

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 Engineer*ing, July 21, 2009, http://www.epistemics.so.uk/Notes/61-0-0.htm

Noy., N.F., & McGuinness, D.L, 2001, *Ontologi 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

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

LAMPIRAN 1

DAFTAR RIWAYAT PENELITI (KETUA)

I. DATA DIRI

1. Nama Lengkap : Herny Februariyanti, ST., M.Cs
2. NIY : YS.2.01.01.035
3. Golongan/Pangkat : IIIC/ Penata Tk.I
4. Jabatan Fungsional : Lektor
5. Tempat, Tgl. Lahir : Blora, 14 Pebruari 1973
6. Jenis Kelamin : Perempuan
7. Alamat Rumah : Jl. Kendeng V/ 12 Semarang
8. Telp / Faks / e-mail : 081 56545909 / - / herny@unisbank.ac.id
9. Alamat Kampus : Jl. Trilomba Juang 1 Semarang
10. Telp / Faks / e-mail : 8311668 / 8443240 / info@unisbank.ac.id

II. RIWAYAT PENELITIAN

No	Judul	Tahun	Keterangan
1	Klastering Dokumen Berita dari Web Menggunakan Algoritma Single Pass Clustering	2011	Ketua
2	Aplikasi Generator Konten untuk Meningkatkan Peringkat Situs pada Halaman Hasil Mesin Pencari.	2010	Anggota
3	Prototipe Mesin Pencari Dokumen Teks	2010	Ketua
4	Aplikasi Pengelolaan Peraturan Daerah Provinsi Jawa Tengah Menggunakan Basisdata XML	2010	Anggota
5	Aplikasi Pengindeks Kata Berbasis Web Pada Dokumen Teks Berbahasa Indonesia Untuk Keperluan Temu Kembali Informasi.	2009	Ketua
6	Hierarchical Agglomerative Clustering untuk Sistem Temu Kembali Dokumen Bahasa Indonesia	2009	Ketua
7	Pengindeks Kata Dokumen Teks dengan Menggunakan Aplikasi Berbasis Web	2009	Ketua

Demikian daftar riwayat penelitian ini dibuat dengan sebenarnya.

Semarang, 15 Juli 2012

Herny Februariyanti, ST., M.Cs

DAFTAR RIWAYAT PENELITI (ANGGOTA)

I. DATA DIRI

1. Nama Lengkap : Eri Zuliarso, Drs., M.Kom
2. NIY : YU.2.10.11.097
3. Golongan/Pangkat : IIIB / Penata Muda Tk II
4. Jabatan Fungsional : Asisten Ahli
5. Tempat, Tgl. Lahir : Temanggung, 23 November 1968
6. Jenis Kelamin : Laki-laki
7. Alamat Rumah : Jl. Pucang Permai III/2, Mranggen, Demak
8. Telp / Faks / e-mail : 085876470885/ ..... / ezuliarso@yahoo.com
9. Alamat Kampus : Jl. Trilomba Juang 1 Semarang
10. Telp / Faks / e-mail : 8311668 / 8443240 / info@unisbank.ac.id

II. RIWAYAT PENELITIAN

No	Judul	Tahun	Keterangan
1	Klastering Dokumen Berita Menggunakan Algoritma Single Pass Clustering	2011	Anggota
2	Web Service Memanfaatkan Layanan Facebook	2010	Ketua
3	Pembuatan Crawling Web	2010	Ketua
4	Prototipe Mesin Pencari Dokumen Teks	2010	Anggota
5	Aplikasi Pengolah Bahasa Alami Untuk Query Basis Data XML Akademik	2009	Ketua

Demikian daftar riwayat penelitian ini dibuat dengan sebenarnya.

Semarang, 15 Juli 2012

Drs. Eri Zuliarso, M.Kom

DAFTAR RIWAYAT PENELITI (ANGGOTA)

I. DATA DIRI

1. Nama Lengkap : Rina Anwaristyati, S.Kom
2. NIY : Y.3.96.12.066
3. Golongan/Pangkat : IIIA/-
4. Jabatan Fungsional : -
5. Tempat, Tgl. Lahir : Kudus / 4 November 1975
6. Jenis Kelamin : Perempuan
7. Alamat Rumah : Jl. Pucang Argo Tengah I/14 Pucanggading Demak
8. Telp / Faks / e-mail : 081325214945/../R1n44nw4R@gmail.com
9. Alamat Kampus : Jl. Trilomba Juang 1 Semarang
10. Telp / Faks / e-mail : 8311668 / 8443240 / info@unisbank.ac.id

II. RIWAYAT PENELITIAN

No.	Judul	Tahun	Keterangan
1	Klastering Dokumen Berita Menggunakan Algoritma Single Pass Clustering	2011	Anggota

Demikian daftar riwayat penelitian ini dibuat dengan sebenarnya.

Semarang, 15 Juli 2012

Rina Anwaristyati, S.Kom

DAFTAR RIWAYAT PENELITI (ANGGOTA)

I. DATA DIRI

1. Nama Lengkap : Moh Sefrian Nugroho
2. NIM : 09.01.55.0081
3. Golongan/Pangkat : -
4. Jabatan Fungsional : -
5. Tempat, Tgl. Lahir : Semarang, 25 Mei 1991
6. Jenis Kelamin : Laki-laki
7. Alamat Rumah : Jl. Lamper Tengah 3 Semarang
8. Telp / Faks / e-mail :08985534755/muhammadsefrian@yahoo.com
9. Alamat Kampus : Jl. Trilomba Juang 1 Semarang
10. Telp / Faks / e-mail : 8311668 / 8443240 / info@unisbank.ac.id

II. RIWAYAT PENELITIAN

No.	Judul	Tahun	Keterangan

Demikian daftar riwayat penelitian ini dibuat dengan sebenarnya.

Semarang, 15 Juli 2012

Moh Sefrian Nugroho

DAFTAR RIWAYAT PENELITI (ANGGOTA)

I. DATA DIRI

1. Nama Lengkap : Berbudhi Rachman Hidayat
2. NIM : 09.01.55.0034
3. Golongan/Pangkat : -
4. Jabatan Fungsional : -
5. Tempat, Tgl. Lahir : Semarang, 27 September 1990
6. Jenis Kelamin : Laki-laki
7. Alamat Rumah :Jl.Griya Prasetya Selatan 3 no127 Semarang
8. Telp / Faks / e-mail : 08985767693 / sasimiyo@gmail.com
9. Alamat Kampus : Jl. Trilomba Juang 1 Semarang
10. Telp / Faks / e-mail : 8311668 / 8443240 / info@unisbank.ac.id

II. RIWAYAT PENELITIAN

No.	Judul	Tahun	Keterangan

Demikian daftar riwayat penelitian ini dibuat dengan sebenarnya.

Semarang, 15 Juli 2012

Berbudhi Rachman Hidayat

LAMPIRAN 2

LOKASI PENELITIAN