SISTEM PEMBANGKIT ANOTASI PADA ARTIKEL BERGAMBAR DENGAN PENDEKATAN KONTEKSTUAL

Diana Purwitasari, Dian Sahputra, Esti Yuniar, Umi Laili Yuhana, Daniel Siahaan
Lab Semantik, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember
Jl. Raya ITS – Gedung Teknik Informatika ITS Sukolilo Surabaya 60111
Telp : (+6231) 5939214, Fax : (+6231) 5913804
Email: diana@if.its.ac.id

ABSTRACT

Development of E-learning sites and their materials make it is necessary to help users finding the desired materials. Context-based search engine will help users for the finding task. However that kind of searching can only be done for learning materials that have been semantically signed or annotated. Annotation is given for the article's content or the article's image within. There are many constraints for manually providing annotations to the learning articles such that automatic metadata or annotation generating method is needed. This paper discusses about annotation generating system with two subsystems: annotation recommender for learning material using contextual analysis and image metadata generator. The methods for contextual analysis are Latent Semantic Analysis (LSA) and WordNet-lexical dictionary usage. Our experimental results showed that subsystems can be used to generate annotation for articles and images in the articles though we have not done combination of two subsystems.

***Keywords**: automatic anotation, ontology, latent semantic analysis, opennlp, text processing*.

1. PENDAHULUAN

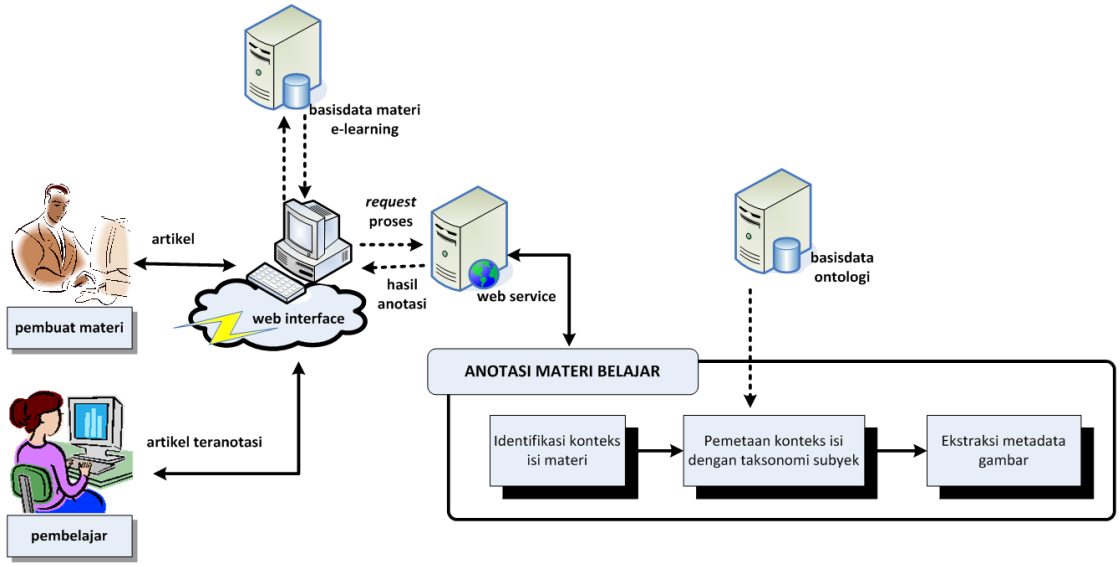
E-learning menjadi tempat tersedianya materi dan tugas pembelajaran seperti latihan soal/ujian. Berdasarkan hasil survei Top 100 e-Learning di tahun 2009 terdapat materi belajar gratis tersedia di Internet yang setidaknya mencapai 200 jam waktu belajar [1]. Dikarenakan pertambahan materi belajar terus meningkat diperlukan cara mempermudah pengguna dalam menemukan artikel berisi materi yang diinginkan. Untuk memperbaiki kualitas hasil pencarian perlu dibuat mesin pencari yang dapat memahami konteks isi dari artikel dengan bahasa manusia (*context-based search engine*). Analisa pemahaman kontekstual dari artikel disebut sebagai interpretasi secara semantik. Namun pencarian hanya bisa dilakukan pada materi belajar yang sudah memiliki penanda semantik/anotasi tertulis. Oleh karena itu metadata materi harus ditambahkan oleh pembuat materi.

Pemberian anotasi pada artikel e-learning akan memudahkan dalam memahami isi artikel karena anotasi dapat berfungsi sebagai penanda topik yang sedang dibahas. Berdasarkan anotasi suatu materi belajar maka pembelajar dapat menemukan artikel-artikel lain dengan konteks subyek yang masih terkait dengan materi belajar tersebut. Label anotasi dapat dihasilkan dari abstraksi isi artikel secara kontekstual dengan teknik berbasis semantik [2] untuk membantu dalam pengorganisasian materi.

Pertambahan jumlah materi juga membuat organisasi materi belajar menjadi sangat dibutuhkan. Beberapa domain pengetahuan bahkan telah memiliki taksonomi subyek untuk materi belajar sebagai hasil konsensus pemikiran para ahli yang dapat digunakan sebagai acuan pengorganisasian materi seperti domain *software engineering* dengan www.swebok.org.

Topik penelitian dalam makalah ini adalah otomatisasi sistem rekomendasi anotasi yang akan menghasilkan daftar subyek-subyek dari taksonomi sebagai kandidat anotasi pada materi belajar. Materi belajar dapat ditulis dengan gabungan teks dan gambar. Metadata terkait dengan gambar memerlukan pendekatan anotasi yang berbeda. Pada makalah ini pembuatan metadata yang masih dilakukan manual [3] akan berubah menjadi otomatis. Tambahan anotasi hasil pemetaan isi materi dengan subyek pada taksonomi membuat pembelajar dapat mencari artikel-artikel yang memiliki konteks materi serupa. Artikel-artikel yang dianggap relevan akan terdaftar berdasarkan urutan belajarnya dengan adanya anotasi berdasarkan taksonomi subyek.

Kontribusi dari penelitian ini adalah metode untuk menghasilkan rekomendasi anotasi materi belajar secara otomatis dengan melakukan pemetaan isi materi dan taksonomi subyek menggunakan pemahaman semantik. Setelah metode tersebut diimplementasikan, pencarian berbasis konteks akan memperhitungkan metadata dari anotasi dalam prosesnya. Penelitian ini bermanfaat bagi pembelajar untuk membantu mencari materi belajar yang konteks isinya sesuai dengan subyek yang diinginkan. Manfaat lebih lanjut terkait dengan daftar materi



Gambar 1. Gambaran Sistem Rekomendasi Anotasi Berbasis Semantik

dalam hasil pencarian yang telah tersusun sesuai dengan urutan subyek belajar karena anotasi pada materi belajar diambil dari taksonomi subyek suatu domain pengetahuan.

2. DESKRIPSI SISTEM PEMBANGKIT ANOTASI

Proses anotasi artikel pada dasarnya dibagi ke dalam 2 bagian utama, yaitu: proses anotasi artikel dan proses anotasi gambar yang ada dalam artikel. Proses dalam sistem ditunjukkan pada Gambar 1.

Langkah-langkah yang akan dilakukan dalam proses rekomendasi anotasi adalah:

1. Pembuat artikel terlebih dahulu mengakses web untuk menulis artikel yang diinginkan.
2. Sistem di client akan melakukan request ke web service untuk memproses anotasi pada artikel dan anotasi pada citra.
3. Untuk proses anotasi artikel dilakukan langkah-langkah berikut:
   a. Proses pencarian istilah-istilah penting dengan metode Latent Semantic Analysis (LSA) [4].
   b. Setelah mendapatkan istilah-istilah penting, selanjutnya dilakukan proses pengukuran kemiripan istilah-istilah yang terpilih dengan topik yang ada di taksonomi SWEBOK.
   c. Topik-topik di taksonomi SWEBOK yang memiliki nilai kemiripan paling tinggi dengan istilah-istilah penting yang terpilih akan dikembalikan ke pembuat artikel sebagai bentuk rekomendasi anotasi. Pembuat artikel akan memilih dan menyetujui rekomendasi yang diberikan oleh sistem dan kemudian metadata hasil akan dikirimkan ke database materi elearning. Selanjutnya jika artikel yang telah teranotasi memiliki citra atau gambar, maka RSS artikel akan dikirimkan ke proses anotasi citra.
4. Untuk proses anotasi citra, hal-hal yang dilakukan antara lain:
   a. Proses pengambilan RSS dan metadata artikel. RSS dan metadata yang ada dalam artikel diekstrak bersamaan dengan gambar yang ada dalam artikel tersebut.
   b. Gambar-gambar yang ada di dalam artikel akan dianotasi dengan metadata yang ada di dalam artikel ditambahkan dengan informasi yang ada di dalam RSS.
   c. Hasil berupa gambar yang telah dianotasi dan kemudian disimpan ke dalam basisdata materi e-learning.

Pada permasalahan pemberian anotasi pada materi belajar diberikan contoh artikel dengan topik Flowcharts. Saat pembuat materi menulis artikel tersebut, penulis memilih topik Flowcharts and structured flowchart sebagai root-nya. Pilihan topik root berdasarkan taksonomi Software Engineering yang ada pada database aplikasi (berdasarkan SWEBOK). Setelah pembuat materi memutuskan untuk menyimpan artikel maka sistem akan bekerja untuk mendapatkan rekomendasi anotasi melalui langkah-langkah yang diadaptasikan dari teknik LSA.

Latent Semantic Analysis (LSA) adalah sebuah teknik matematika/statistik untuk menggali dan menyimpulkan hubungan kontekstual dari kata-kata dalam sebuah wacana [4]. Ide yang melandasi adalah melakukan agregat dari semua konteks kata yang diberikan, baik yang ada ataupun tidak dalam menyediakan batasan untuk menentukan kesamaan arti dari kata terhadap set kata yang lainnya. Pada makalah ini bahasan tentang LSA tidak dijelaskan secara rinci. Bahasan mendetil dapat dilihat langsung

pada referensi tentang LSA [4]. Hal yang dibahasa pada makalah ini adalah penerapan LSA pada konteks permasalahan.

Pada LSA, hubungan antara kata dan dokumen akan direpresentasikan ke dalam k-dimensional space yang sama untuk dibandingkan. Ini yang biasa disebut dengan latent semantic space. LSA menggunakan Singular Value Decomposition (SVD) yaitu teknik analisa faktor pada matriks untuk memproses kata-kata yang ada dalam dokumen. LSA biasanya diterapkan pada koleksi dokumen untuk mendapatkan latent semantic space. Hasil dari LSA adalah konteks isi yang ada di koleksi dokumen. Pengambilan frekuensi kemunculan term ini bisa dilakukan menggunakan fungsi yang ada pada Oracle Text [5]. Setelah konteks isi dari materi belajar diidentifikasi, langkah selanjutnya akan menghitung kemiripan kata-kata yang terpilih dengan topik-topik pada taksonomi subyek menggunakan WordNet.

Untuk artikel yang berisi gambar diperlukan metadata khusus untuk gambar tersebut. Metadata akan digunakan pada pencarian. Pemberian metadata dilakukan dengan ekstraksi isi dari artikel yang disebut dengan tokenisasi. Di dalam topik penelitian ini akan dilakukan proses tokenisasi dengan memanfaatkan library openNLP [7] seperti proses tokenisasi. Pada bagian ini akan diberikan contoh langkah-langkah yang dilakukan dalam proses tokenisasi untuk memacah struktur sebuah kalimat. Hasil tokenisasi akan digunakan untuk proses berikutnya.

3. IMPLEMENTASI SISTEM PEMBANGKIT ANOTASI

Sistem pembangkit anotasi akan dibangun dari dua subsistem yaitu anotasi teks dan anotasi teks dengan gambar. Untuk selanjutnya subsistem pertama akan disebut dengan sistem pembangkit anotasi berdasarkan konteks. Kemudian subsistem kedua akan disebut sebagai sistem pembangkit anotasi pada artikel bergambar.

3.1 Gambaran umum tentang implementasi subsistem

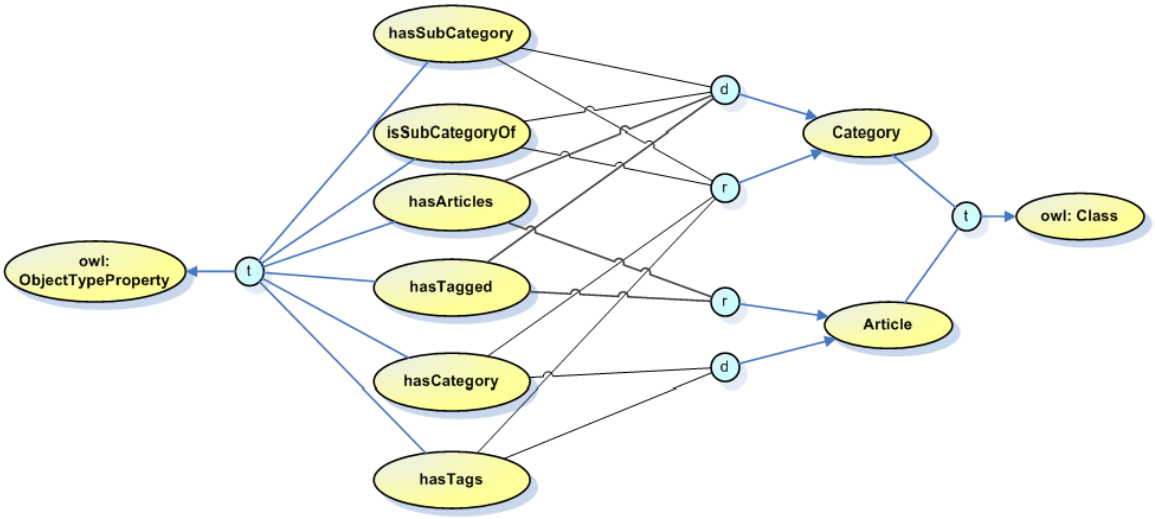
Desain proses pada sistem pembangkit anotasi berdasarkan konteks terdiri dari empat proses utama yakni ekstraksi fitur term, rekomendasi anotasi dengan metode Latent Semantic Analysis (LSA), rekomendasi anotasi dengan metode WordNet, dan pemberian anotasi pada artikel. Untuk melakukan anotasi berdasarkan konteks pada suatu artikel teks sistem harus mengetahui topik – topik yang menjadi konteks bahasan dalam artikel terlebih dulu. LSA dapat memetakan sekumpulan artikel ke dalam ruang topik sedemikian hingga artikel – artikel dengan konteks bahasan yang sama akan terlihat dekat dalam visualisasi pemetaan. Untuk mempermudah visualisasi LSA dapat diaproksimasi menjadi ruang topik dua dimensi. Namun jika kata – kata kunci yang menjadi penyusun suatu topik tidak selalu sama antar setiap artikel maka digunakan bantuan WordNet sebagai kamus leksikal untuk mencari kata sinonim. Bahasa menjadi batasan karena WordNet yang digunakan adalah kamus leksikal untuk kata – kata Inggris. Aktivasi kesemua proses dilakukan langsung oleh pengguna sehingga tidak diperlukan proses khusus oleh seorang administrator. Namun sistem harus sudah dilengkapi dengan data koleksi artikel teks dan data ontologi untuk standar ruang topik.

Sedangkan implementasi sistem pembangkit anotasi pada artikel bergambar memiliki lima proses utama yakni melakukan konfigurasi sistem pembangkit metadata, melakukan proses pembangkitan metadata RDF, menghentikan proses pembangkitan metadata, melakukan pengamatan metadata yang telah dibangkitkan, dan mencari citra berdasar metadata RDF. Proses – proses tersebut diperuntukkan bagi administrator dan pengguna. Administrator bertugas untuk melakukan penyiapan anotasi pada sekumpulan artikel bergambar sedemikian hingga pengguna dapat melakukan pencarian artikel berdasarkan informasi pada gambar.

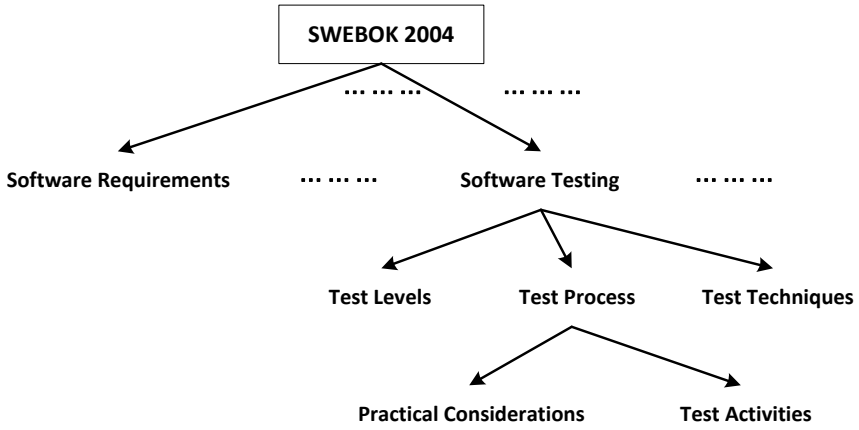
3.2 Implementasi subsistem pembangkit anotasi berdasarkan konteks

Desain sistem pembangkit anotasi terdiri dari empat proses utama. Proses pertama ekstraksi fitur term terdiri dari dua langkah yaitu pendataan kata dalam artikel yang dibantu dengan library Oracle Text [5] dan kemudian kata – kata tersebut akan diberi bobot TF-IDF [8]. Isi artikel yang akan diproses disimpan dalam database Oracle kemudian dilakukan proses pengindeksan yang menghasilkan daftar kata dalam artikel tersebut beserta jumlah kemunculannya. Pengindeksan dilakukan untuk tiap paragraf sebagai sub artikel dan menganggapnya sebagai artikel baru. Jadi topik dari suatu artikel akan dikenali berdasarkan topik dari setiap sub artikel. Proses pembobotan TF-IDF dilakukan untuk menentukan kata – kata yang dianggap penting dalam artikel dari daftar kata yang diperoleh dari pendataan kata. TF dari suatu kata merupakan frekuensi kemunculan kata tersebut pada suatu artikel. IDF suatu kata adalah nilai logaritmik dari frekuensi kemunculan kata dalam artikel di suatu koleksi. Nilai bobot TF-IDF suatu kata merupakan hasil perkalian nilai TF dan IDF untuk kata tersebut. Bobot nilai TF-IDF untuk semua daftar kata kemudian diurutkan dari yang paling tinggi hingga yang paling rendah. Dari bobot TF-IDF yang telah urut diambil 25 % bobot teratas. Pemilihan nilai 25% ini diambil setelah dilakukan uji coba pemilihan bobot.

Proses kedua rekomendasi anotasi dengan metode LSA [4] menggunakan matriks pemetaan



Gambar 2. Model Data Anotasi Ontologi pada Sistem Pembangkit Anotasi Berdasarkan Konteks



Gambar 3. Contoh Hubungan Kategori Berdasarkan Taksonomi SWEBOK

kata dan artikel dengan nilai setiap elemen merupakan nilai bobot TF-IDF kata dalam artikel. Matriks tersebut merupakan ruang fitur dari koleksi yang menggunakan frekuensi kemunculan kata sebagai dasar pemetaannya. LSA digunakan untuk menghitung nilai kedekatan konteks bahasan dari kategori dengan suatu artikel sehingga kategori tersebut dapat dijadikan rekomendasi anotasi untuk artikel. Untuk itu diperlukan daftar kategori beserta relasi antar kategori. Relasi diperlukan untuk mencari kandidat anotasi yang memiliki konteks bahasan sama. Pada implementasi sistem digunakan daftar kategori yang telah tersusun dalam taksonomi standar untuk domain bahasan pemrograman perangkat lunak berdasarkan Software Engineering Book of Knowledge (SWEBOK, www.swebok.org).

Hubungan antar kategori dapat dijelaskan dengan class RDF pada Gambar 2. Sebagai contoh dari kategori – kategori yang ada pada Gambar 3, hubungan antar kategori dapat dijelaskan sebagai berikut:

- Test Process isSubCategoryOf Software Testing
- Software Testing hasSubcategory Test Levels
- Software Testing hasSubcategory Test Techniques
- Test Process hasSubcategory Practical Considerations
- Test Process hasSubcategory Test Activities

Sedemikian hingga *parent* dari kategori Test Process adalah kategori Software Testing; *children* dari kategori Test Proces adalah kategori Practical Considerations dan Test Activities; *siblings* dari kategori Test Proces adalah kategori Test Levels dan Test Techniques.

Untuk mendapatkan rekomendasi anotasi dengan metode LSA perlu didapatkan nilai scales, term vector, dan doc vector terlebih dahulu. Kemudian langkah selanjutnya adalah menghitung nilai kedekatan antara kategori relasi dengan artikel. Kategori sebagai kandidat anotasi dianggap sebagai query. Sehingga akan terbentuk query vector yang merupakan hasil penjumlahan term vector dari kata-kata yang ditemukan pada query. Selanjutnya mencari skor artikel sebagai kriteria kedekatan konteks bahasan artikel dengan kategori yang merupakan hasil dot product dari scales, query

vector, dan doc vector. Pada implementasi penghitungan nilai kedekatan antara kategori relasi dan artikel dengan LSA digunakan library lingpipe.jar (http://alias-i.com/lingpipe/) untuk menghitung Singular Value Decompotition (SVD).

Proses ketiga rekomendasi anotasi dengan metode WordNet akan dilakukan apabila metode LSA tidak dapat menghasilkan rekomendasi. Pada metode ini penentuan suatu kategori dijadikan rekomendasi anotasi atau tidak didasarkan pada jarak kemiripan antara kategori relasi dengan kata-kata penting [9]. Pada implementasi penghitungan nilai kedekatan dengan WordNet digunakan library WordNet.Net [10].

Proses keempat pemberian anotasi pada artikel dilakukan setelah pengguna memilih kategori yang akan dianotasikan dari rekomendasi anotasi yang dihasilkan sistem. Hasil dari proses ini adalah artikel yang telah tesimpan dalam sistem yang dilengkapi dengan anotasi.

Pada sistem yang diimplementasikan proses pendapatan rekomendasi anotasi dengan metode LSA dan atau Wordnet akan dijalankan saat menambah artikel baru ataupun mengedit artikel lama.

3.3 Implementasi subsistem pembangkit anotasi pada artikel bergambar

Terdapat dua proses utama dalam implementasi sistem pembangkit anotasi pada artikel bergambar. Proses pertama adalah melakukan konfigurasi sistem pembangkit metadata. Konfigurasi yang dilakukan adalah menentukan alamat RSS dari situs artikel bergambar, set koneksi internet untuk pengambilan artikel dan memastikan jeda waktu pengecekan RSS untuk mengetahui artikel bergambar terbaru. Situs artikel BBC Sport dipilih untuk uji coba dalam implementasi sistem.

Proses kedua terkait dengan pembangkitan metadata RDF dari artikel bergambar didahului dengan mendapatkan isi artikel sesuai dengan daftar artikel terbaru dari RSS. Sistem akan melakukan permintaan untuk mendapatkan script XML dari RSS yang sedang aktif setiap satu jam sekali. Untuk mengambil isi artikel dimulai dengan melakukan cek pada item RSS dan mendapatkan link yang dibutuhkan. Kemudian sistem akan melihat jikalau artikel tersebut memiliki gambar dengan bantuan library HTMLParser [11].

Pengecekan keberadaan gambar dalam artikel dilakukan dengan memperhatikan tag html yang dimiliki oleh artikel. Sistem tidak akan mengunduh artikel tidak bergambar. Keluaran proses adalah informasi alamat url gambar dan url isi artikel. Selanjutnya dilakukan pemrosesan pada isi artikel untuk mengekstrak kata – kata penting terkait dengan gambar yang ada dalam artikel. Sistem menghasilkan file RDF sebagai metadata artikel bergambar dengan bantuan library Jena [12].

Artikel bergambar ditentukan sebagai artikel yang disediakan oleh situs berita olahraga karena gambar menyatakan aktifitas tertentu terkait dengan bahasan dalam artikel. Aktifitas memiliki satu atau lebih pelaku berbentuk subject dan action dilakukan oleh subject. Informasi yang lebih mendetail dalam gambar antara lain event terjadinya aktifitas place dan times yang menunjukkan waktu dan tempat pengambilan. Ekstraksi data dari artikel untuk mengisi RDF class dilakukan dengan bantuan library OpenNLP khususnya fungsi EnglishTreeBankParser [7]. Fungsi OpenNLP tersebut dapat memisahkan frasa-frasa kalimat Bahasa Inggris. Implementasi fungsi EnglishTreeBankParser mengambil struktur kalimat yang merupakan Noun Phrase dan Verb Phrase. Noun Phrase akan diproyeksikan sebagai instance dari subject sementara verb dalam Verb Phrase diproyeksikan menjadi action. Model yang disediakan oleh OpenNLP hanyalah person, organization, location, date, time, percentage, dan money. OpenNLP tidak dapat menghasilkan RDF class sesuai untuk class event sehingga diperlukan suatu model pembelajaran.

Kemudian setelah sistem pembangkit anotasi pada artikel bergambar menghasilkan keluaran file RDF sebagai metadata, maka pengguna dapat mencari gambar berdasarkan metadata RDF berupa halaman web pencarian berisi combo box filter subject.

3.3 Implementasi sistem pembangkit anotasi

Implementasi sistem pembangkit anotasi merupakan kombinasi dari implementasi dua subsistem terdahulu. Proses penggabungan kedua subsistem tersebut akan menjadi pekerjaan selanjutnya. Sebelum dilakukan penggabungan subsistem maka pada setiap subsistem akan dilakukan pengujian. Hasil dari pengujian disebutkan pada bahasan berikut.

4. EVALUASI DAN HASIL UJI COBA

Skenario pengujian telah ditetapkan untuk evaluasi dari implementasi sistem pembangkit anotasi berdasarkan konteks dan sistem pembangkit anotasi pada artikel bergambar.

4.1 Pengujian dalam sistem pembangkit anotasi berdasarkan konteks

Data yang digunakan pada uji coba adalah 30 artikel mengenai Software Engineering didownload dari www.wikipedia.org periode Februari - Juli 2010. Artikel-artikel tersebut akan disimpan dalam aplikasi pembangkit anotasi pada materi e-learning dengan pendekatan semantik untuk mendapatkan hasil rekomendasi anotasi. Terdapat beberapa skenario uji coba yang akan dilakukan yaitu pemilihan kata-kata penting dengan pembobotan TF-

IDF dan perhitungan presisi hasil rekomendasi anotasi.

Skenario pengujian pertama tentang pemilihan kata – kata akan dibahas sebagai berikut. Panjang artikel dalam koleksi terdiri dari sekitar 130 kata diperoleh melalui pengindeksan Oracle dengan 20 kata berbeda diantaranya yang muncul lebih dari satu kali. Pengindeksan artikel hanya memperhitungkan teks dan tidak mengikutkan gambar dalam prosesnya. Kata – kata tersebut kemudian dihitung nilai bobotnya dengan skema TF-IDF. Penghitungan bobot menjadi salah satu kriteria filter untuk mengurangi jumlah kata dalam daftar kata indeks. Dicoba tiga kemungkinan nilai ambang sebagai penyeleksian kata yaitu 20%, 25%, dan 30%. Dari daftar kata indeks pada suatu artikel dilihat 20% bobot tertinggi dan kemudian bobot yang paling bawah dalam 20% tersebut dijadikan batas bawah suatu kata dianggap sebagai kata penting. Cara yang sama dilakukan untuk bobot 25% dan 30%.

Hasil penyeleksian 20% mendapatkan jumlah kata penting yang terlalu sedikit dengan artikel memiliki hanya satu kata penting. Sedangkan 30% memberikan jumlah kata penting yang terlalu banyak dengan 9 artikel yang jumlah kata pentingnya sama dengan jumlah kata muncul lebih dari satu kali sehingga penyeleksian kata seolah tidak dilakukan. Oleh karena itu penyeleksian dengan nilai 25% dipilih sebagai penentu batas minimal bobot kata penting. Sehingga kata penting yang tersisa dari daftar kata indeks pada suatu artikel berkurang menjadi tersisa 40% kata.

Skenario pengujian kedua tentang precision dan recall [13] hasil rekomendasi anotasi akan membandingkan tingkat kebenaran pemberian anotasi otomatis LSA dan WordNet dengan anotasi manual oleh ahli. Penilaian precision dan recall memerlukan kunci jawaban kategori relasi yang seharusnya menjadi rekomendasi dengan dipilih secara manual setelah membaca isi artikel data uji coba.

Rumus untuk menghitung *precision* dan *recall* adalah sebagai berikut:

$$precision = \frac{jml\ rekomendasi\ sesuai}{jml\ rekomendasi\ didapat}\ x\ 100\% \quad (1)$$

$$recall = \frac{jml\ rekomendasi\ sesuai}{jml\ rekomendasi\ seharusnya}\ x\ 100\% \quad (2)$$

Hasil sistem diharapkan memiliki nilai *precision* yang tinggi namun nilai *recall* diharapkan tidak terlalu berbeda jauh dengan nilai *precision*. Nilai *recall* dapat dengan mudah diperoleh yakni dengan cara sistem akan menghasilkan rekomendasi semua kategori yang ada dalam SWEBOK. Namun hal tersebut tidak diinginkan karena anotasi kategori artikel menjadi tidak jelas maknanya. Oleh karena itu digunakan kriteria pengukuran *F measure* yang biasanya digunakan dengan mempertimbangkan nilai *precision* dan *recall* [13].

Rumus untuk menghitung *F measure* adalah sebagai berikut:

$$F\ measure = \frac{2\ x\ precision\ x\ recall}{precision + recall} \quad (3)$$

Nilai F measure untuk rekomendasi LSA, WordNet dan kombinasi keduanya adalah 64.98%, 53.39% dan 73.32%. Sistem dengan rekomendasi LSA lebih baik (64.98%) dibandingkan dengan sistem rekomendasi WordNet (53.39%). Hal itu disebabkan LSA akan memetakan setiap artikel ke ruang topik sesuai dengan kategori SWEBOK. Sedangkan rekomendasi dengan WordNet hanya memperhatikan artikel dengan judul yang diperkirakan bersinonim dengan kategori. Perkiraan sinonim dilakukan dengan mempertimbangkan kamus leksikal WordNet. Apabila sistem tidak dapat menghasilkan rekomendasi kategori sebagai anotasi dengan metode LSA maka sistem akan mencari rekomendasi kategori dengan WordNet. Perhitungan nilai kedekatan judul artikel yang akan dicari anotasinya dan kategori dalam taksonomi SWEBOK dengan metode WordNet pasti menghasilkan nilai kedekatan > 0. Sehingga untuk menentukan kategori yang paling tepat dijadikan rekomendasi adalah dengan mengambil 30% nilai kedekatan tertinggi dari daftar kategori yang dihasilkan dengan metode WordNet. Kombinasi pemakaian 2 metode tersebut dalam menghasilkan rekomendasi memperbaiki tingkat keakuratan sistem sebanyak 13% dengan nilai F measure untuk keduanya adalah 73.32%.

Pengujian pada sistem pembangkit anotasi berdasarkan konteks telah menunjukkan bahwa anotasi berupa kategori dalam taksonomi subyek suatu domain pengetahuan dapat dihasilkan. Analisa anotasi berdasarkan konteks dilakukan dengan memetakan artikel ke ruang topik kategori dengan metode LSA. Untuk mengurangi ruang fitur kata dalam pemetaan maka telah dilakukan pengujian bobot kata yang paling sesuai. Namun apabila cara tersebut tidak dapat dilakukan maka dicari kategori terdekat dengan memperhatikan sinonim judul yang dianggap sebagai representasi bahasan topik dari artikel.

4.2 Pengujian dalam sistem pembangkit anotasi pada artikel bergambar

Skenario pengujian pertama yang dilakukan adalah memastikan proses – proses dalam sistem pembangkit anotasi pada artikel bergambar bisa dilakukan. Proses pertama adalah melakukan konfigurasi sistem pembangkit metadata. Proses kedua adalah menghasilkan metadata RDF dari artikel bergambar yang didahului dengan

mendapatkan isi artikel sesuai daftar artikel terbaru dari RSS.

Evaluasi dilakukan pada koleksi artikel diambil dari RSS untuk pengujian tingkat keakuratan metadata yang dihasilkan oleh sistem. Sejumlah 45 artikel terambil dari RSS yang didapatkan pada tanggal 1 Agustus 2010 pukul 13.00 dari BBC Sport. Kategori sport diambil dengan alasan gambar yang ada pada artikel jenis kategori tersebut biasanya merepresentasikan isi artikel. Dari 45 artikel terambil dengan topik bahasan umumnya mengenai kategori football, rugby union, dan rugby league ternyata hanya 32 artikel yang memiliki gambar. Kemudian ahli akan mendefinisikan metadata untuk koleksi 32 artikel tersebut secara manual. Evaluasi dilakukan dengan membandingkan metadata yang dihasilkan oleh sistem serta metadata yang didefinisikan melalui pengamatan manual.

Sebagai contoh terdapat suatu artikel sebagai berikut:

Berdasarkan pengamatan manual didefinisikan metadata untuk artikel tersebut adalah: subject: beckham; action: watch.

Kemudian sistem pembangkit anotasi pada artikel bergambar menghasilkan metadata sbb:

subject: beckham; action: watch.

Dikarenakan terdapat dua metadata dengan setiap metadata memiliki nilai yang sama antara hasil diperoleh sistem dan hasil pengamatan manual maka tingkat relevansi metadata adalah 100%.

Evaluasi dilakukan dengan mengamati tingkat relevansi untuk kesemua 32 artikel yang diunduh. Jenis metadata yang akan dievaluasi adalah subject, action, dan event. Hasil evaluasi ditunjukkan pada Tabel 1.

Hasil uji coba menunjukkan bahwa sistem dengan bantuan OpenNLP akan lebih mudah dalam mengenali event. Pada kasus ini OpenNLP untuk event perlu dilakukan pembelajaran terlebih dahulu untuk mengenali event yang benar. Tingkat akurasi terendah diperoleh apabila sistem diminta untuk menghasilkan metadata jenis action. Hal tersebut disebabkan karena OpenNLP mengalami kesulitan saat mengenali kata kerja. Secara overall relevansi metadata masih rendah jika dibandingkan metadata yang hanya berisi subyek dikarenakan pengenalan subyek jauh lebih mudah.

Meskipun berdasarkan evaluasi uji coba masih banyak hal perlu diperbaiki terutama dalam menghasilkan metadata class action, namun sistem dapat menghasilkan anotasi untuk artikel bergambar.

5. SIMPULAN

Hasil penelitian menunjukkan bahwa peneliti berhasil mengimplementasikan sistem pembangkit anotasi pada artikel dan gambar yang ada pada suatu artikel dengan membuat dua sub sistem [14, 15]. Sistem anotasi gambar yang dibuat masih terbatas untuk gambar yang menyatakan aktifitas atau aksi dari suatu subyek di sebuah artikel. Selain itu anotasi artikel dan gambar yang berhasil dibuat sementara masih ditujukan untuk artikel berbahasa Inggris. Meskipun kedua sub sistem masih berjalan sendiri-sendiri, hasil uji coba menunjukkan bahwa dengan memberikan anotasi atau metadata pada artikel dan gambar, artikel dan gambar dapat ditemukan dengan mesin pencari berbasis semantik. Penggabungan kedua sub sistem akan dilakukan pada penelitian berikutnya.

6. DAFTAR PUSTAKA

[1] Garner, E., 2009, Manage, Click, Learn!, url:www.managetrainlearn.com

[2] Chu, H., Chen, M., & Chen, Y., 2009, A Semantic-based Approach to Content Abstraction and Annotation for Content Management, Expert Systems with Applications, (36) 2360–2376.

[3] Awaludin, M., Siahaan, D.O., & Yuhana, U.L., 2009, Sistem Navigasi dan Pencarian Berbasis Konteks pada Konten E-Learning dengan Teknologi Web Semantik, Tugas Akhir di Teknik Informatika Institut Teknologi Sepuluh Nopember.

[4] Landauer, T.K., Foltz P.W., & Laham D., 1998, An Introduction to Latent Semantic Analysis, Discourse Process, 25(2-3), 259–84.

[5] Shea, C., Faisal, M., Ford, R., Lin, W., & Matsuda, Y., 2008, Oracle Text Application Developer's Guide, 11g Release 1 (11.1).

[6] Miller, G. A., 1995, WordNet: A Lexical Database for English, Communications of The ACM 38(11) 39–41.

[7] Baldridge, J., & Morton, T., 2010, OpenNLP, Diakses pada 1 Juli 2010 dengan url http://opennlp.sourceforge.net

[8] Salton, G., & Buckley, C., 1988, Term-Weighting Approaches in Automatic Text Retrieval, Information Processing and Management, 24(5), 513-523.

[9] Dao, T. N., & Simpson, T., 2008, Measuring Similarity Between Sentences, Diakses pada 10 April 2010 dengan url http://opensvn.csie.org/WordNetDotNet/trunk/Projects/Thanh/Paper/WordNetDotNet_Semantic_Similarity.pdf.

[10] Simpson, T., & Malcolm, C., 2005, WordNet.Net. Diakses pada 10 Mei 2010, dengan url http://opensource.ebswift.com/WordNet.Net

[11] Oswald, D., Raha, S., Macfarlane, I., & Walters, D., 2006, HTML Parser, Diakses 1 Juli 2010, http://htmlparser.sourceforge.net

[12] HP Labs Semantic Web Research, 2009, Jena - A Semantic Web Framework for Java, Diakses 1 Juli 2010, url http://jena.sourceforge.net
[13] Manning, C.D., Raghavan, P., & Schütze, H., 2009, An Introduction to Information Retrieval: ch8 – Evaluation in information retrieval, Cambridge University Press.
[14] Purwitasari, D., Yuniar, E., Yuhana, U.L., & Siahaan, D.O., (dipublikasikan pada 2011), Ontology-based Annotation Recommender for Learning Material Using Contextual Analysis, Proc. of the IETEC'11 Conf., Kuala Lumpur, Malaysia.
[15] Yuhana, U.L., Sahputra, D. Purwitasari, D., & Siahaan, D.O., 2010, Pengembangan Sistem Pembangkit Metadata Citra untuk Citra pada Situs Olah Raga, Seminar Sistem Informasi Indonesia 2010 (SESINDO 2010), Surabaya