TINJAUAN SIMILARITAS SEMANTIK DALAM PEMELIHARAAN ONTOLOGI PADA *PEER-TO-PEER* (*P2P*)

Lintang Yuniar Banowosari[1)], I Wayan Simri Wicaksana[2)]

[1,2] Universitas Gunadarma, Jl. Margonda Raya no.100, Depok 16424, Indonesia
{lintang, iwayan}@staff.gunadarma.ac.id

ABSTRAK

Isu tentang sumber informasi informasi yang terdapat pada Internet yaitu : massive (sangat besar), terdistribusi, dinamis, dan open, menyebabkan keragaman semantik [11]. Untuk mengatasi keragaman tersebut beberapa pendekatan telah dilakukan, salah satunya adalah dengan menggunakan pendekatan semantik yang digabungkan dengan P2P.

P2P [10] memungkinkan terjadinya pembentukan komunitas yang memiliki kesamaan interest. Dengan terbangunnya group (Semantic Overlay Network-SON) ini maka perbedaan semantik dapat dikurangi. Hal ini sangat penting untuk aktifitas pertukaran informasi pada suatu domain, misalnya tumbuhan.

Pada model P2P, ontologi kerap diasumsikan sudah terbentuk sebelumnya. Tetapi dalam menghadapi lingkungan yang dinamis pada P2P, ontologi yang sudah terbentuk kerap tidak lagi memenuhi konsep dari anggota komunitas. Sehingga diperlukan sebuah pendekatan khusus untuk pemeliharaan ontologi pada lingkungan P2P.

Pemeliharaan ontologi dengan melihat konsep di provider peer, akan memerlukan proses mapping dan merging dalam mencapai aligment. Sebelum melakukan proses mapping dan merging perhitungan similaritas [7,8] adalah sangat penting. Setiap ontologi dapat direpresentasikan dalam sebuah hirarki label terminologi.

Pada paper ini kami akan menguji beberapa model pendekatan label similarity dengan berbasiskan string calculation, latent semantic, dan taksonomi seperti WordNet [15]. Kami akan menguji beberapa schema dan common ontology dari berbagai domain seperti tanaman, bisnis, pendidikan.

Kata Kunci: *latent semantic*, *ontology*, pemeliharaan, *P2P*, semantik, similaritas, *WordNet*

1. PENDAHULUAN

1.1. Keragaman Informasi

Internet dan Web merupakan sumber informasi yang semakin lama semakin besar, hal ini memunculkan masalah dalam beberapa isu tentang sumber informasi yaitu : massive (sangat besar), terdistribusi, dinamis, dan open.

Menurut Sheth [11] terdapat dua kelompok keragaman yaitu: keragaman informasi dan keragaman sistem. Keragaman informasi menyebabkan munculnya perbedaan dari sistem informasi. Perbedaan bisa terjadi pada tinggkatl sintaktis, struktur, dan semantik. Untuk mengatasi keragaman tersebut beberapa pendekatan telah dilakukan, salah satunya adalah dengan menggunakan pendekatan interoperabilitas semantik yang digabungkan dengan *P2P*.

P2P memungkinkan terjadinya pembentukan komunitas yang memiliki kesamaan interest. Dengan terbangunnya *group* ini maka perbedaan semantik dapat dikurangi. Model ini kerap disebut dengan *Semantic Overlay Network (SON*). Tetapi pendekatan ini belum memadai sehingga tetap memerlukan jembatan dengan memanfaatkan pendekatan mediasi semantik yang didukung oleh ontologi.

Penggunaan ontologi dan *P2P* telah semakin berkembang dalam beberapa tahun terakhir ini. Karena managemen pengetahuan dan konten dalam *P2P* arsitektur lebih mudah dilakukan dibandingkan dengan sistem terbuka penuh.

Pada model *P2P*, ontologi kerap diasumsikan sudah terbentuk sebelumnya. Tetapi dalam menghadapi lingkungan yang dinamis pada *P2P*, ontologi yang sudah terbentuk kerap tidak lagi memenuhi konsep dari anggota komunitas. Sehingga diperoleh sebuah pendekatan khusus untuk pemeliharaan ontologi pada lingkungan *P2P*.

Pemeliharaan ontologi dengan melihat konsep di *provider peer*, akan memerlukan proses *mapping* dan *merging* dalam mencapai *aligment*. Sebelum melakukan proses *mapping* dan *merging* perhitungan similaritas [7] adalah sangat penting.

1.2. Arsitektur *P2P*

Pengertian *P2P* sangat beragam, Milojick [10] mengumpulkan beberapa definisi, yang dapat disimpulkan dalam karakter yang dimiliki oleh *P2P* sebagai berikut : berbagi, pertukaran langsung, mengorganisasi sendiri dan independen, node dapat menjadi server atau client, pengalamatan dan sistem koneksi yang independen.

Arsitektur P2P yang dibahas akan menggunakan hybrid model dengan super peer (SP). SP akan menyimpan common ontology (CO) sebagai acuan atau pivot untuk kegiatan pertukaran informasi. Selama pertukaran informasi akan terjadi agreement / mapping antara sebagian common ontology dengan sebagian ontologi lokal di peer yang memiliki sumber informasi (provider peer / PP). Semakin tinggi tingkat agreement maka tingkat akurasi pertukaran informasi semakin baik. Untuk meningkatkan tingkat agreement salah satunya adalah dengan memelihara common ontology.

Model pertukaran informasi pada P2P seperti di atas adalah dengan menggunakan pendekatan mediasi semantik. Pada mediasi semantik akan diperlukan beberapa komponen sebagai berikut:

- Lokal Konteks, terdiri dari :
  - Data lokal yang terdapat pada PP dan yang akan digunakan secara bersama oleh komunitas, dapat dalam bentuk data relasional atau XML/RDF/OWL.
  - Skema eksport di PP akan merepresentasikan lokal data untuk publik. Skema eksport ini kerap juga disebut dengan ontologi lokal.
  - Wrapper adalah sarana untuk menjembatani antara skema ekspor ke/dari lokal data. Wrapper bukan saja digunakan untuk merubah format data, tetapi juga representasi data, query, dan respon.
- Komunitas Konteks
  - Common ontology (CO) adalah merupakan representasi konsep dari komunitas. CO memegang peranan sangat penting untuk referensi dari anggota komunitas. CO diletakkan di SP.
- Pemetaan Konteks
  - Agreement atau pemetaan adalah merupakan hal penting untuk dapat terjadinya pertukaran informasi antara peer anggota dari komunitas. Agreement merupakan pemetaan dari skema eksport ke common ontology dan disimpan pada PP. Agreement akan terdiri dari subset secara agreement unit, dan dinyatakan dalam model :
    AU=<LO,CO,LC> (1)
    dimana :
    AU : agreement unit
    LO : ontologi lokal
    CO : common ontology
    LC : pemetaan local ke common ontology

Dari tiga konteks di atas jelas common ontology memegang peranan sangat penting untuk tingkat keberhasilan pertukaran informasi dalam sebuah komunitas P2P.

1.3. Konsep Ontologi

Pengertian ontologi sangat beragam, dari definisi Benjamins [1]: “Sebuah Ontologi merupakan definisi dari pengertian dasar dan

relasi vokabulari dari sebuah area sebagaimana aturan dari kombinasi istilah dan relasi untuk mendefinisikan vokabulari".

Gruber [2] memberikan definisi yang banyak diacu, yaitu "Ontologi merupakan sebuah spesifikasi eksplisit dari konseptualisme". Guarino dan Giaretta pada 1995 mengumpulkan tujuh definisi yang berkoresponden dengan *syntactic* dan *semantic.* Pada 1997, Borst melakukan modifikasi dari definisi Gruber dengan mengatakan "Sebuah ontologi adalah spesifikasi formal dari sebuah konseptual yang diterima (share)".

Sebuah ontologi dijelaskan dengan menggunakan notasi dari konsep, *instances*, relasi, fungsi, dan aksiom [2].

• Konsep dapat pula merupakan penjelasan dari tugas, fungsi, aksi, strategi, dan sebagainya.

• Relasi 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 ... x Cn, contoh : subclass-of dan connected-to.

• Fungsi 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, contoh : Mother-of.

• Aksiom digunakan memodelkan sebuah sentence yang selalu benar.

• Instances adalah digunakan untuk merepresentasikan elemen.

Tujuan ontologi adalah menangkap pengetahuan dari sebuah domain dan disajikan secara generik dan memberikan kesamaan pandangan dan pemahaman dari domain tersebut.

2. PEMELIHARAAN ONTOLOGI

Pemeliharaan ontologi dapat melalui berbagai pendekatan, pendekatan secara umum adalah :

- mapping, dimana dipetakan satu ontologi ke ontologi lainnya,
- merging, dimana digabungkan dua atau lebih ontologi menjadi sebuah ontologi
- alignment dimana penyesuaian ontologi karena ada perubahan atau penyesuaian knowledge dan konsep.

Perhitungan similaritas diperlukan saat proses mapping.

2.1. Tinjauan Similaritas

Pemeliharaan ontologi dengan melihat konsep di *provider peer*, akan memerlukan proses mapping dan merging dalam mencapai *aligment*. Sebelum melakukan proses *mapping* dan *merging* perhitungan similaritas adalah sangat penting. Setiap ontologi dapat direpresentasikan dalam sebuah hirarki label terminologi.

Perhitungan *semantic similarity* adalah merupakan proses yang memerlukan keterlibatan beberapa disiplin ilmu, seperti bahasa, komputer, matematika logik dan domain yang bersangkutan. Langkah awal perhitungan kesamaan *semantic* adalah mengacu kepada kesamaan terminological atau kerap kali disebut label. Terminologi yang dimaksud dapat meliputi class, property hingga instances. Menurut Euzenat [4], pendekatan terminological ada yang berdasarkan string based dan *language based*. Pada paper ini akan ditinjau pendekatan untuk *languange based* dengan menggunakan *lexicons* (seperti *WordNet*) dan *string-based.*

2.1.1 Metode Penghitungan Persamaan Similaritas.

2.1.1.2 *WordNet*

WordNet merupakan sebuah leksikal database elektronik. WordNet dikembangkan untuk bahasa Inggris oleh Universitas Princeton di Amerika.

WordNet adalah sistem referensi leksikal online yang rancangannya terinspirasi oleh teori psikolinguistik dari memori leksikal manusia. Kata benda, kata kerja, kata sifat dan kata keterangan dalam bahasa Inggris diorganisir menjadi himpunan sinonim, dimana masing masing merepresentasikan satu konsep leksikal. Relasi yang berbeda menghubungkan himpunan sinonim.

Pada WordNet beberapa informasi dapat dicari seperti persamaan kata, lawan kata, arti kata (glossary), singkatan bahkan juga sampai kepada beberapa hal yang penting untuk sistem informasi seperti:

- taksonomi, 'matahari' adalah bagian (subClass) dari 'tata surya'
- agregasi, 'genteng' adalah bagian (part of) dari 'rumah'
- kemiripan, [anjing,kuncing] > [anjing, pohon]

Metode kesamaan semantik perhitungan pada WordNet dibagi dalam dua kelompok besar pendekatan [9], yaitu *path lenght* dan *information content*. *Path lenght* secara sederhana menghitung jumlah node atau relasi yang menghubungkan antar node dalam taksonomi. Jarak yang lebih pendek antara dua konsep,bararti memiliki kesamaan lebih tinggi. *Pathlenght* memberikan keuntungan dengan tidak bergantung pada statistik *corpus* dan tidak terpengaruh dengan penyebaran kata. Tetapi memiliki kelemahan dalam taksonomi yang memiliki jarak yang *uniform*/sama. Beberapa contoh pendekatan dengan *path lenght* adalah Leacock-Chodorow, Resnik, Wu-Palmer. Pada paper ini, pendekatan Wu Palmer adalah salah satu model yang diuji, persamaanya seperti pada formula 1

Wu-Palmer:

$$sim_{WUP} = \max\left[\frac{2\,xdepth(LCS(a,b))}{lenght(a,b)+2\,xdepth(LCS(a,b))}\right] \quad (1)$$

dimana length(a, b) adalah jumlah panjang path antara a dan b; depth(LCS(a, b)) adalah jumlah panjang path dari konsep umum dari a dan b ke root. *Information content* sebuah node adalah -log dari jumlah semua kemungkinan (dihitung berdasarkan frekuensi corpus) dari semua kata yang memiliki *synset*. *Synset* merupakan sebuah set sinonim dari sebuah konsep yang sama dipasangkan dengan penjelasannya seperti gloseri dari *synset*. Dengan kata lain jika p(x) adalah probability dari sebuah *instace* dari x, maka *information content* dari x adalah −log p(x). Salah satu pendekatan yang popular adalah Jiang Conrath pada persamaan 2

Jiang-Conrath:

$$sim_{JCN} = \max [IC(a) + IC(b) - 2xIC(LCS(a, b))] \quad (2)$$

dimana IC(a) dan IC(b) adalah *information content* dari node a sebagai −log jumlah dari semua probabilitas (dihitung dari frekuensi corpus) untuk semua kata pada *synset*; IC(LCS(a, b)) adalah *information content* pada sebuah node konsep umum atau bersama dari a dan b.

2.1.1.2 *String-based*

Metode yang berbasis string menggunakan struktur dari string itu sendiri (sebagai satu urutan dari huruf). Metode berbasis string biasanya akan menemukan dan menganggap `Match` dan `match` adalah kelas yang serupa, tapi tidak untuk `alignment.`

Terdapat lebih banyak cara untuk membandingkan string daripada cara yang dapat dilakukan dalam melihat string (sebagai satu urutan huruf yang pasti, satu urutan huruf yang salah, sebuah himpunan kata)..

2.1.1.2.1. *Levenshtein Distance*

Levenshtein atau *edit* distance [4] adalah suatu pendekatan yang berbasis string yang digunakan untuk menghitung jarak/perbedaan string. *Edit distance* mendefinisikan string dengan sembarang panjangnya, dan dapat menghitung perbedaan antara 2 string, di mana perhitungan perbedaan tidak hanya ketika string mempunyai perbedaan karakter tetapi salah satunya mempunyai karakter tertentu, sedangkan string yang lain tidak. Definisi formal adalah sebagai berikut:

r(a, b) = 0 if a = b. Let r(a, b) = 1,otherwise. (3)

Misalkan diberikan 2 string s dan t dengan panjang d dan m. Kita akan mengisi array d (n+1)x(m+1) dengan bilangan bulat positif (integer) sehingga elemen pojok kanan yang paling rendah d(n+1,m+1) akan memindahkan/menyediakan nilai yang dibutuhkan Levenshtein distance L(s,t). Definisi masukan dari d adalah rekursif. Himpunan pertama d(i,0) =i, i = 0,1,...,n dan d(0,j) = j, j = 0,1,...,m.

2.1.1.2.2. *Euclidian N-Gram Distance*

Jarak Euclidian [4] adalah jarak yang "biasa" antara dua (2) titik yang salah satunya dapat diukur dengan penggaris, yang dapat dibuktikan melalui aplikasi dari teorema Pythagoras. Dengan menggunakan formula tersebut sebagai jarak, ruang Euclidian akan menjadi ruang metrik atau metrik pythagoras.

Jarak Euclidian antara dua (2) titik

$P = (p_1, p_2, \ldots, p_n)$ dan $Q = (q_1, q_2, \ldots, q_n)$, dalam n-ruang Euclidian didefinisikan sebagai berikut:

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (4)$$

Distance dua-dimensi sebagai berikut: untuk dua (2) titik 2D: $P = (p_x, p_y)$ dan $Q = (q_x, q_y)$, jarak dihitung sebagai :

$$\sqrt{(p_x - q_x)^2 + (p_y - q_y)^2} \quad (5)$$

3. HASIL PERCOBAAN DAN DISKUSI

3.1 Metode Uji coba.

Tujuan dari percobaan adalah untuk membandingkan beberapa pendekatan dari:

- WordNet dengan menggunakan Wu-Palmer formula dari pendekatan path length
- WordNet dengan menggunakan Jiang Conrath formula dari pendekatan information content
- String-based dengan menggunakan Levenshtein Distance
- String-based dengan menggunakan Euclidian N-Gram Distance

Pengujian akan berdasarkan perbandingan terhadap evaluasi dari ekspert dengan melihat faktor Recall, Precession dan F-measure.

Beberapa hal yang dipersiapkan untuk pengujian adalah :

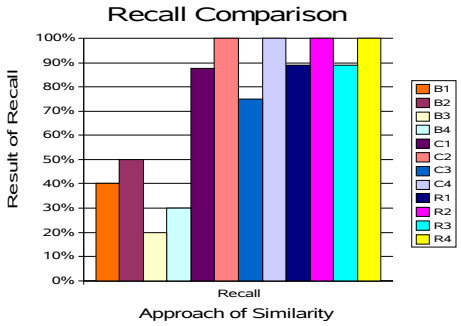
- Mencari atau mengembangkan tool untuk menghitung similaritas berdasarkan pendekatan di atas. Pada percobaan ini digunakan tool on-line yang telah tersedia, untuk WordNet digunakan http://marimba.d.umn.edu/cgibin/*similarity*. cgi, untuk string-based yang menggunakan Levenshtein digunakan http://www.cut-the-knot.org/do_you_know/Strings.shtml , serta yang menggunakan Euclidian N-Gram distance digunakan http://www.josef-willenborg.de/java/NGram/NGramApplet.html.
- Menentukan domain dan konsep yang akan diuji. Pada percobaan ini dilakukan untuk tiga domain yang mencakup transportasi, publikasi buku dan bisnis. Domain ini diambil dari paper yang telah memiliki hasil pengujian *similarity* berdasarkan ekspert pada domain yang bersangkutan. Untuk transportasi mengacu kepada [3], publikasi buku mengacu kepada [6] dan bisnis mengacu kepada [5].

Perhitungan *similarity* akan mengikuti proses sebagai berikut:

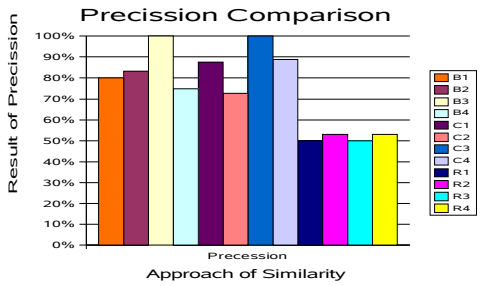
- Menghitung semua kombinasi konsep antara dua sumber dalam satu domain berdasarkan pada keempat pendekatan di atas.
- Memfilter hasil perhitungan dengan memberikan nilai threshold, tujuannya adalah untuk mempermudah dalam analisis. Penentuan nilai threshold dilakukan dengan cara try-error untuk mendapatkan nilai optimal, dimulai dengan initial nilai dari 0.7 ke 1,0. Ini disebut dengan tabel hasil perhitungan ∂.
- Membuat tabel perhitungan dari ekspert atau disebut tabel referensi, ini disebut β.
- Membandingkan hasil perhitungan terhadap referensi ini adalah Δ.
- Kemudian menghitung nilai Recall (Recall =(Δ/β)), Precession (Precession = (Δ/∂)) dan F-Measure (Fmeasure = 2/((1/Recall) + (1/Precession))).
- Hasil perhitungan akan ditampilkan pada grafik utk dievaluasi mana yang memiliki nilai terbaik.

3.2 Hasil dan diskusi

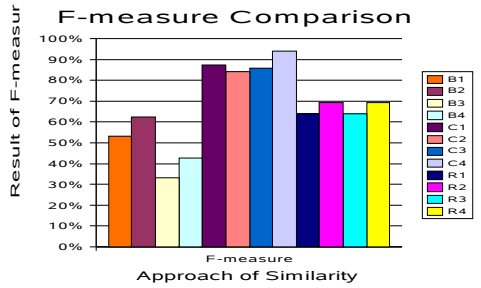
Hasil dari eksperimen ditampilkan pada grafik gambar 1, 2, 3.



Gambar 1 Perbandingan Recall



Gambar 2 Perbandingan Precision



Gambar 3 Perbandingan F-measure

Keterangan gambar 1,2,3:
B1:Transport. domw/WordNet-WUP;B2:Transport.dom. w/WordNet-JCN B3:Transport.dom.String-basedLevenshteinDis.;B4:Transport.dom.String-basedEuclid.Dis.
C1:Book dom.w/WordNet-WUP; C2:Book dom.w/WordNet-JCN C3:Book dom.String-basedLevenshtein Dis.;C4:Book dom.String-based Euclid.Dis.
R1:Bussiness dom.w/WordNet-WUP;R2:Transport. dom.w/WordNet-JCN
R2:Bussiness dom. String-basedLevenshtein Dis;R3:Bussiness. dom. String-based Euclid.Dis.

Dari ketiga gambar tabular di atas, maka didapatkan hasil sebagai berikut :

- *Recall*, *WordNet* dengan Jean Conrath memberikan hasil terbaik pada tiga domain, berarti pendekatan ini mampu menghitung semua similaritas sesuai dengan ekspert.
- *Precession*, semua hasil memiliki hasil yang relatif sama. Walau *edit-distance* menunjukkan hasil sedikit relatif lebih baik.
- *F-measure* sebagai *total performance*, WordNet dengan Wu-Palmer memiliki kecenderungan lebih baik dibandingkan pendekatan yang lain.

Sehingga secara umum dikatakan bahwa dari percobaan yang telah di- lakukan terhadap tiga domain, urutan pendekatan yang terbaik adalah Wu-Palmer, Jean-Conrath, *String based* dengan *Euclidian N Gram Distance*, dan yang terakhir adalah Levenshtein Distance.

4. PENUTUP

Pada paper ini telah dilakukan pengujian terhadap beberapa pendekatan untuk menghitung semantik similaritas. Hasil yang menarik dari percobaan ini adalah dari empat pendekatan tidak memberikan sebuah pendekatan yang sangat menonjol. Sehingga diperlukan penelitian dengan keragaman class yang lebih banyak untuk memberikan hasil yang lebih baik dalam menentukan pilihan pendekatan kesamaan semantik yang akan digunakan.

Pada evaluasi kami, WordNet dengan WUP memberikan kecenderungan nilai hasil yang lebih baik dibandingkan pendekatan lain, tetapi hal ini belum dapat ditarik menjadi kesimpulan akhir.

Mengacu kepada hasil yang dicapai, kami merencanakan untuk mengembangkan penelitian lebih mendalam dengan memperbanyak jumlah class, kedalaman taksonomi skema dan keragaman domain.

5. DAFTAR PUSTAKA

[1] Guarino N., 1998, *"Formal Ontology in Information Systems.",* Proceedings of FOIS'98, Trento, Italy, 6-8 June. Amsterdam, IOS Press, pp. 3-15.

[2] Gruber T.R, 1993, *"A translation approach to portable ontologies"*, Knowledge Acquisition, 5(2):199-220

[3] Yaser Bishr. *Semantic Aspects of Interoperable GIS*. PhD thesis, Wageningen Agricultural University, Netherland, 1997.

[4] Jerome Euzenat, Thanh Le Bach, Jesus Barasa, and etc. D2.2.3: State of the art on ontology alignment. Technical Report IST-2004-507482, knowledgeweb, 2 August 2004.

[5]Erhard Rahm and Philip A. Bernstein. A Survey of Approaches to Automatic Schema Matching. *The VLDB Journal*, 10:334–250, 2001.

[6]Huiyong Xiao, Isabel F Cruz, and Feihong Hsu. Semantic Mappings for the Integration of XML and RDF Sources. In *Proc. of IIWEB-2004*, 30 August 2004.

[7]Lintang Yuniar B., I Wayan Simri W., 2005, "*Pemeliharaan Common Ontology pada P2P dengan Voting dan Representasi"*, Prosiding Seminar Nasional Teknologi Informasi (SNTI) 2005, Universitas Tarumanegara, Jakarta.

[8]Lintang Yuniar B., I Wayan Simri W., 2006, "*Pemeliharaan Ontology pada P2P Berbasis Voting dan Similaritas"*, Prosiding Seminar Nasional SMART 2006, Universitas Gadjah Mada, Jakarta.

[9]Michelizzi J., 2005, "*Similarity and other current activities*", akses Juni 2005 http://www.d.umn.edu/~tpederse/Group04/jm-slides-sep-9.pdf

[10]Milojick D., etc., 2002, "Peer-to-Peer Computing*"*

[11]Sheth A.P, 1998,*"Changing Focus On Interoperability In Information Systems: From System, Syntax, Structure, To Semantics,*" MITRE, Dec 3rd

[12]Wicaksana, 2006, Desertasi Doktor : "*A Peer to Peer (P2P) Based Semantic Agreement Approach for Spatial Information Interoperability"*, Universitas Gunadarma.

[13] Wicaksana, Lintang Y. Banowosari, Lily Wulandari, Setia Wirawan, 2005, " *Pentingnya Peranan Bahasa dalam Interoperabilitas Informasi berbasiskan Komputer karena Keragaman Semantik"*, Prosiding Seminar Ilmiah Nasional (PESAT 2005),Universitas Gunadarma,Jakarta, halaman S9-S16.

[14]WordNet homepage, akses Januari 2005, http://WordNet.princeton.edu