

Membandingkan Pendekatan *Latent Semantic* terhadap WordNet untuk *Semantic Similarity*

I Wayan Simri Wicaksana*

*Universitas Gunadarma

E-mail: iwayan@staff.gunadarma.ac.id

ABSTRAK

Pertukaran informasi antar berbagai sumber di Internet yang semakin otonomi, dinamis dan bebas akan menimbulkan permasalahan tersendiri. Pemahaman sebuah konsep antara berbagai sumber informasi dapat memiliki perbedaan, seperti sebuah konsep yang sama dapat memiliki arti berbeda (contoh bank sebagai institusi keuangan atau tepian sungai), atau konsep yang berbeda dapat memiliki arti yang sama (contoh zip dan post code). Sebagai contoh kita mencari data perusahaan persewaan mobil, untuk menjelaskan konsep/class/field/colom dari jenis mobil dapat menggunakan konsep 'car', 'automobile', atau 'transportation'. Untuk membedakan konsep tersebut bagi manusia adalah 'relatif' mudah tetapi bagi mesin akan menghadapi kesulitan. Ada berbagai pendekatan yang telah dikemukakan pada berbagai riset untuk mengukur kesamaan semantik.

Pada paper ini kami akan membandingkan bagaimana menghitung kesamaan dari konsep secara semantik pada *Latent Semantic* terhadap pendekatan WordNet. Pertama kami akan menguraikan secara ringkas pendekatan dari setiap penghitungan kesamaan semantik. Kemudian kami akan mengujikan pada berbagai konsep pada domain yang berbeda untuk melihat kemampuan dalam mengukur kesamaan semantik. Referensi pengujian hasil pengukuran akan dibandingkan dengan pengukuran kesamaan konsep berdasarkan ekspert atau manusia.

Hasil dari riset kami akan memberikan kontribusi untuk berbagai bidang seperti untuk mesin pencari di Internet, proses mapping, query rewriting, interoperabilitas, dan sebagainya. Kelanjutan dari riset kami adalah untuk menggabungkan dengan penelitian kami yang lain dalam mendukung mapping pada interoperabilitas pada peer-to-peer sebagai pencarian sumber dan query.

Kata Kunci : label matching, latent semantic, semantic similarity, WordNet.

1. Pendahuluan

Interoperabilitas informasi pada masa Internet tidak saja memberikan dampak positif, tetapi juga mengantar kepada berbagai masalah baru. Salah satu permasalahan baru adalah pada keragaman pada sintatik, skematik dan semantik.

Keragaman semantik semakin besar dikarenakan semakin banyaknya pihak yang dapat berpartisipasi dalam pertukaran informasi. Dimana setiap pihak akan memiliki konsep, kepentingan ataupun pemahaman akan 'sesuatu' yang berbeda. Sebagai contoh mencari informasi *transportation*, pada sebuah sumber dengan pemahaman sarana transportasi, sumber lain adalah kendaraan untuk transportasi, sumber lain adalah bisnis dibidang transportasi. Permasalahan yang timbul adalah bagaimana untuk mengukur kesamaan atau perbedaan konsep antar sumber yang bersangkutan.

Pendekatan pengukuran kesamaan semantik (*semantic similarity*) dapat dilakukan

dengan berbagai pendekatan, seperti dengan menggunakan latent semantic atau WordNet.

1.1. Latar Belakang

Semantic similarity adalah sebuah masalah pada hubungan semantik. Hubungan semantik merupakan pendekatan untuk mengetahui bagaimana hubungan dua konsep dalam penggunaan dan relasinya, walaupun beberapa persamaan hanya mempertimbangkan hubungan *IS-A* (*hypernymy / hyponymy*).

Relasi antara konsep adalah tidak selalu simetris, jika dua konsep adalah sama, berarti juga memiliki relasi, tetapi kalau berhubungan belum tentu berarti sama. Problem ini yang kerap menyulitkan dalam perhitungan kesamaan semantik. Sehingga timbul beberapa pendekatan untuk penyempurnaan seperti penggunaan internal dan ekstranal struktur.

1.2. Pembagian Paper

Paper ini akan membandingkan pendekatan dari *Latent Semantic* dari sumber dua

domain terhadap *Semantic Similarity* dengan dua model perhitungan *path length* dan *information content*. Untuk pengujian akan dilihat dari domain transportasi, buku dan bisnis. Pada bagian 1 akan menguraikan latar belakang yang mencakup definisi dan permasalahan utama. Bagian ke-dua menjelaskan pendekatan semantic similarity dengan menggunakan *latent semantic* dan WordNet. Perbandingan kedua metode akan didiskusikan pada bagian 3. Bagian ke empat merupakan kesimpulan.

2. Semantic Similarity

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 [2], pendekatan terminological ada yang berdasarkan string based dan language based. Pada paper ini akan ditinjau pendekatan untuk language based dengan menggunakan lexicons (seperti WordNet) dan latent semantic.

2.1. WordNet

WordNet adalah sebuah database network semantik untuk bahasa Inggris yang dikembangkan di Princeton University (<http://wordnet.princeton.edu/>). Beberapa versi dalam bahasa lain juga telah dikembangkan seperti EuroNet.

Bagian dasar dari WordNet adalah *synset*. *Synset* merupakan sebuah set sinonim dari sebuah konsep yang sama dipasangkan dengan penjelasannya seperti glosari dari *synset*. *Synset* dihubungkan dengan berbagai bentuk relasi seperti *hyponymy* (adalah jenis dari), *meronymy* (adalah bagian dari), *antonymy* (adalah lawan dari) dan sebagainya.

Metode kesamaan semantik perhitungan pada WordNet dibagi dalam dua kelompok besar pendekatan [7], yaitu *path length* dan *information content*. *Path length* secara sederhana menghitung jumlah node atau relasi yang menghubungkan antar node dalam taksonomi. Jarak yang lebih pendek antara dua konsep, berarti memiliki kesamaan lebih tinggi. *Path*

length 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 length* 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:

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

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 akata yang memiliki *synset*. Dengan kata lain jika $p(x)$ adalah probability dari sebuah instance dari x , maka *information content* dari x adalah $-\log p(x)$. Salah satu pendekatan yang populer 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.

2.2. Latent Semantic

Latent Semantic Analysis (LSA) [4] adalah sebuah teori dan metode untuk mengekstrak dan merepresentasikan konteks yang digunakan sebagai sebuah arti kata dengan memanfaatkan komputasi statistik untuk sejumlah *corpus* yang besar dari teks. Ide yang mendasari 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. LSA yang memadai akan merefleksikan pengetahuan manusia yang dinyatakan dalam berbagai cara.

Secara sederhana proses dari LSA adalah sebagai berikut :

- Merepresentasikan teks dalam matrik, dimana baris menunjukkan kata yang unik dan kolom adalah dokumen yang bersangkutan. Setiap *cell* akan menunjukkan jumlah/frekuensi kata pada setiap dokumen. Dari langkah ini akan membuat matriks $\{X\}$.
- Selanjutnya LSA melakukan *singular value decomposition* (SVD) terhadap matriks di atas. Pada SVD, matriks akan didekomposisi kedalam produk dari tiga matriks. Satu komponen matriks menjelaskan orignal dari baris entiti sebagai vektor dari nilai ortogonal, yang lain menjelaskan kolom orignal dan yang ketiga sebagai sebuah matrik diagonal yang terdiri dari nilai sekala terhadap ke tiga komponen matrik. Proses ini melakukan dekomposisi matriks $\{X\} = \{W\} \{S\} \{P\}$.

3. Pengujian

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
- Latent semantic dengan menggunakan *corpus* dari General Reading up to 1st year collage
- Latent semantic dengan menggunakan *corpus* dari Encyclopedia

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

3.1. Persiapan Pengujian

Beberapa hal yang dipersiapkan untuk pengujian adalah :

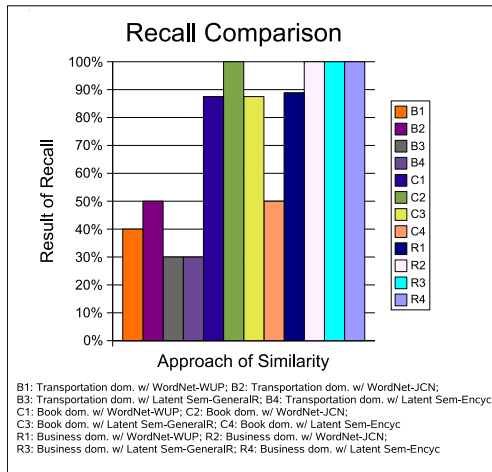
- Mencari atau mengembangkan tool untuk menghitung similaritas berdasarkan pendekatan di atas. Pada percobaan ini digunakan tool on-line yang telah tersedia, untuk WordNet digunakan <http://marimba.d.umn.edu/cgi-bin/similarity.cgi>, untuk

latent semantic digunakan <http://lsa.colorado.edu/>.

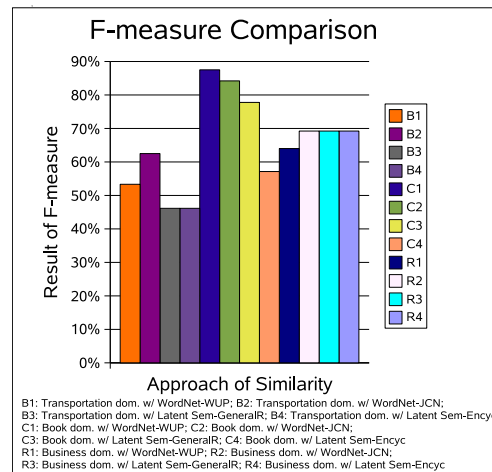
- Menentukan domain dan konsep yang akan diuji. Pada percobaan ini dilakukan untuk tiga domain yang mencakup transportasi, publikasi buku dan bisnis. Domain ini diambil dari paper yang telah memiliki hasil pengujian similarity berdasarkan ekspert pada domain yang bersangkutan. Untuk transportasi mengacu kepada [1], publikasi buku mengacu kepada [8] dan bisnis mengacu kepada [6].
- Perhitungan similarity akan mengikuti proses sebagai berikut:
 - Menhitung 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 (ζ).
 - Membuat tabel perhitungan dari ekspert atau disebut tabel referensi, ini disebut β
 - Membandingkan hasil perhitungan terhadap referensi ini adalah Δ
 - Kemudian menghitung nilai Recall ($Recall = (\Delta/\beta)$), Precision ($Precision = (\Delta/\zeta)$) dan F-Measure ($Fmeasure = 2/((1/Recall) + (1/Precision))$).
- Hasil perhitungan akan ditampilkan pada grafik utk dievaluasi mana yang memiliki nilai terbaik.

3.2. Hasil dan Diskusi

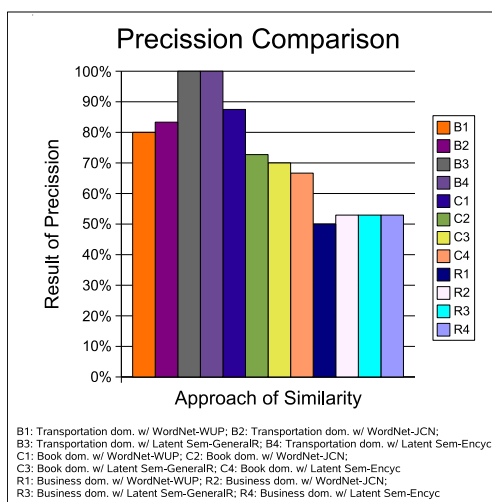
Hasil dari eksperimen ditampilkan pada grafik 1, 2, 3.



Gambar. 1. Hasil Recall



Gambar. 3. Hasil F-Measure



Gambar. 2. Hasil Precession

Dari ketiga grafik di atas, maka didapatkan hasil sebagai berikut :

- Pada Recall, WordNet dengan Jean Conrath memberikan hasil terbaik pada tiga domain, berarti pendekatan ini mampu menghitung semua similaritas sesuai dengan ekspert, walaupun apakah semua yang dihitung sesuai.
- Pada Precession, tidak ditemukan pendekatan yang menonjol, walau WordNet dengan Wu-Palmer relatif sedikit lebih baik dalam memberikan akurasi perhitungan similaritas.
- Pada 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, Latent Semantic dengan General Reading dan terakhir Latent Semantic dengan Encyclopedia. Walaupun dari beberapa riset [7], [5] di WordNet dikatakan bahwa perhitungan kesamaan semantik dengan information content (Jean-Conrath) lebih baik hasilnya dibandingkan terhadap path length (Wu-Palmer). Perbedaan ini bisa terjadi dikarenakan menggunakan versi WordNet yang berbeda antara percobaan pada paper ini dengan paper lainnya. Pada percobaan di paper ini menggunakan versi 2.1 sedangkan riset lainnya menggunakan versi dibawah 2.0. Selain itu perbedaan domain yang digunakan juga berbeda, pada paper lain hanya digunakan 1 domain saja, sedangkan pada paper ini menggunakan 3 domain.

4. Penutup

Pada paper ini telah dilakukan pengujian terhadap beberapa pendekatan untuk menghitung semantik similaritas. Hasil yang menarik dari percobaan ini adalah pendekatan WordNet dengan *path length* lebih baik dibandingkan *information content*, hal ini berbeda dari dua referensi sebelumnya. Perbedaan ini bisa disebabkan perbedaan versi WordNet ataupun juga perbedaan domain dan konsep yang diuji.

Kontribusi dari paper ini adalah dapat dimanfaatkan untuk pengembangan aplikasi semantic web, perawatan ontology, pemetaan konsep dalam memilih pendekatan perhitungan kesamaan semantik yang sesuai.

Langkah ke depan, kami akan mengujikan kepada domain dan kosep yang lebih luas, sehingga dapat menghasilkan kesimpulan yang lebih generik. Disisi lain adalah mengembangkan aplikasi yang bersifat semi otomatis untuk memasukkan multi entri word secara sekaligus dengan format output yang beragam. Hal ini untuk mempercepat proses perhitungan dan pengolahan data.

DAFTAR PUSTAKA

- [1] Yaser Bishr. *Semantic Aspects of Interoperable GIS*. PhD thesis, Wageningen Agricultural University, Netherland, 1997.
- [2] 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.
- [3] Thomas Hofmann. Probabilistic latent semantic analysis. In Kathryn B. Laskey and Henri Prade, editors, *UAI*, pages 289–296. Morgan Kaufmann, 1999.
- [4] K. Thomas Landauer, Peter W. Foltz, and Darell Laham. An Intoduction to Latent Semantic Analysis. In *Discourse Processes*, number 25, pages 259–284, 1998.
- [5] Dekang Lin. An information-theoretic definition of similarity. In Jude W. Shavlik, editor, *ICML*, pages 296–304. Morgan Kaufmann, 1998.
- [6] Erhard Rahm and Philip A. Bernstein. A Survey of Approaches to Automatic Schema Matching. *The VLDB Journal*, 10:334–250, 2001.
- [7] Martin Warin. Using WordNet and Semantic Similarity to Disambiguate an Ontology. http://ling16.ling.su.se:8080/PubDB/doc_repository/warin2004usingwordnet.pdf, 2004.
- [8] 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.