Makalah Nomor: KNSI-469

QUERY REWRITING BERBASIS SEMANTIK MENGGUNAKAN WORDNET DAN LCh PADA SEARCH ENGINE GOOGLE

Ahmad M. Thantawi[1], I Wayan Simri Wicaksana[2], Lily Wulandari[3]

[1] Teknik Informatika, [2] Teknologi Informasi, [3] Teknik Informatika
[1] Fakultas Teknik, UPI YAI, [2] Pascasarjana, [3] Fakultas Teknologi Industri, Universitas Gunadarma
[1] Jl. Diponegoro 74 Jakarta, [2,3] Jl. Margonda Raya No. 100 Pondok Cina Depok
[1] thantawi@yai.ac.id, {[2] iwayan, [3] lily}@staff.gunadarma.ac.id

Abstrak

Saat ini sumber dari informasi semakin bertambah secara cepat pada kurun waktu terakhir ini, terlebih dengan makin berkembangnya teknologi internet. Besarnya jumlah sumber informasi juga melahirkan keragaman dari sumber informasi tersebut. Pengumpulan data dengan memanfaatkan Internet seperti adanya fasilitas Google adalah dengan cara memasukkan kata kunci. Mesin pencari tidak selalu memberikan informasi yang akurat. Kekurangan ini biasanya disebabkan oleh dua masalah utama. Pertama, mesin pencari tidak mampu menemukan pola dari dokumen relevan. Kedua, pengguna tidak menyatakan permintaannya dengan benar. Pada artikel ini menjelaskan teknik bahwa penulisan ulang permintaan (*query*) dengan menggunakan semantik similarity yaitu leacock & chodrow pada saat pencarian informasi di mesin pencari dapat memberikan hasil informasi yang lebih tepat.

Kata kunci : *query rewriting, semantik, wordnet, leacock & chodrow*

1. Pendahuluan

Memasuki era globalisasi dan teknologi informasi yang berkembang pesat, kebutuhan akan informasi yang cepat dan akurat semakin besar. Sumber dari informasi semakin bertambah secara cepat pada kurun waktu terakhir ini, apalagi dengan makin berkembangnya teknologi internet. Besarnya jumlah sumber informasi juga melahirkan keragaman dari sumber informasi tersebut. Untuk menemukan informasi yang kita butuhkan, kita dapat menggunakan sumber informasi yang ada, salah satunya adalah dengan menggunakan bantuan mesin pencari (*search-engine*) di internet, salah satunya adalah mesin pencari Google.

Mesin pencari merupakan alat yang sangat berguna untuk mencari informasi di alam dunia maya. Mesin pencari memiliki kemampuan untuk mencari informasi dari dokumen-dokumen yang tidak terstruktur. Kemampuan ini sangat berguna ketika kita ingin mencari suatu informasi dari dokumen-dokumen yang masing-masing memiliki struktur yang berbeda.

Walaupun memiliki manfaat yang menjanjikan, mesin pencari tidak selalu memberikan informasi yang akurat. Kekurangan ini biasanya disebabkan oleh dua masalah utama. Pertama, mesin pencari tidak mampu menemukan pola dari dokumen relevan. Kedua, pengguna tidak menyatakan permintaannya dengan benar, misalnya dengan menggunakan kalimat yang redundan. Masalah pertama dapat diselesaikan dengan memperbaiki teknologi mesin pencari, sehingga mesin pencari dapat mengenali pola dokumen-dokumen relevan. Masalah kedua dapat diselesaikan dengan algoritma pencarian [1].

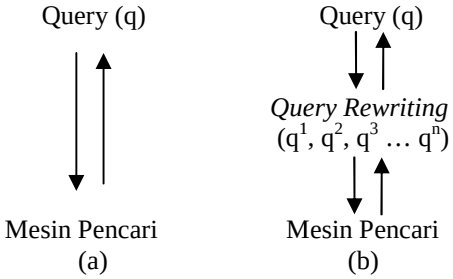
Dari uraian diatas memberikan gambaran bahwa penulisan ulang permintaan (*query*) pada saat pencarian informasi di mesin pencari dapat memberikan hasil informasi yang lebih baik, maka timbul pertanyaan bagaimana pendekatan yang digunakan dalam rangka penulisan ulang suatu permintaan (*query*) informasi pada mesin pencari agar didapatkan informasi yang lebih baik dan metode atau pendekatan apa yang digunakan dalam mengukur tingkat kesamaan semantik dari konsep permintaan (*query*) yang ada ?

2. Pendekatan

2.1 Ilustrasi *Query Rewriting*

Query Rewriting adalah sebuah proses penulisan ulang suatu kueri asli ke kueri yang baru dengan menyesuaikan konsep atau terminalogi yang digunakan di masing-masing sumber data

dalam sistem yang terintegrasi, seperti yang tergambar di bawah ini :



Gambar 1 Ilustrasi *query rewriting*

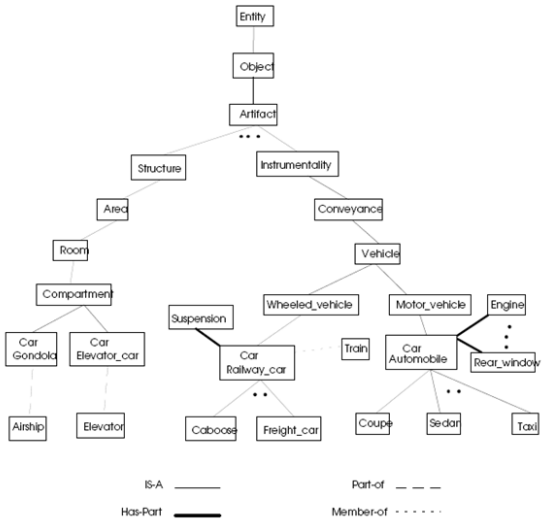
Sebagai contoh adalah jika ada permintaan (*query*) untuk mencari informasi "Jumlah Mobil di Jakarta", pada saat permintaan itu dikirimkan ke sumber data seperti yang tergambar pada gambar 1 point a, maka hanya akan didapatkan hasil yang berkaitan dengan permintaan tadi, tetapi jika dilakukan proses *query rewriting* diharapkan tidak hanya informasi "Jumlah Mobil di Jakarta" yang dihasilkan tetapi informasi "Jumlah kendaraan di Jakarta", "Jumlah kendaraan roda empat di Jakarta", "Jumlah Truck di Jakarta" dapat dihasilkan seperti yang tergambar pada gambar 1 point b.

Pengaksesan informasi berdasarkan kata kunci kadang-kadang tidak memuaskan, dikarenakan penggunaan kata kunci yang dikonversi menjadi sebuah pertanyaan (*query*) tidak dapat menghasilkan jawaban (*response*) yang diharapkan. Sebagai ilustrasi, misalkan kita mengirimkan sebuah permintaan informasi ke beberapa institusi (asumsikan semua institusi memiliki sumber informasi elektronik yang dapat diakses oleh publik) di Depok untuk mencari berapa jumlah tenaga kerja (*employee*). Misalkan permintaan informasi dikirimkan ke berbagai institusi seperti perusahaan swasta, kantor pemerintah, lembaga pendidikan. Untuk kantor pemerintah tenaga kerja diistilahkan dengan pegawai (*employee*), beberapa pabrik menggunakan istilah buruh (*labor*), sementara perusahaan swasta diistilahkan pekerja (*worker*), dan di universitas memakai kata dosen (*lecture*). Kalau kita hanya mengacu kepada query "berapa jumlah tenaga kerja di Depok / *how many employee at Depok*". Maka informasi yang bisa dijawab berdasarkan pendekatan keyword adalah hanya untuk kantor pemerintahan, sedangkan dari institusi lainnya akan memberikan informasi dengan nilai nol. Walaupun kita tahu bahwa antara *employee*, *labor*, *worker*, *lecture* adalah hal yang sama [3].

2.2 Wordnet

WordNet adalah sebuah database network semantik untuk bahasa Inggris yang dikembangkan di Princenton University. WordNet berisi informasi tentang kata benda, katakerja, kata sifat dan kata keterangan. Ia mengorganisir konsep yang terkait ke dalam kumpulan sinonim atau synsets. Masing-Masing synset dapat dipertimbangkan mewakili suatu konsep atau makna/pengertian kata. Sebagai contoh: {car, auto, automobile, machine, motorcar} adalah suatu synset yang menghadirkan makna/pengertian: kendaraan bermotor yang beroda 4; pada umumnya yang didorong oleh suatu mesin pembakaran di bagian dalam. Pada masing-masing synset menghadirkan suatu konsep atau makna/pengertian kata [3]**.**

Wordnet telah menjadi suatu sumber daya yang populer untuk mengidentifikasi hubungan jaringan dan taksonomi antar konsep. Pada Gambar 2 menunjukkan bagaimana WordNet melakukan ekstrak dari suatu konsep [2].



Gambar 2. Ekstrak WordNet untuk konsep 'Car'

Synset dihubungkan dengan berbagai bentuk relasi seperti hypernym (adalah jenis dari), meronymy (adalah bagian dari), antonymy (adalah lawan dari) dan sebagainya. Jika sebuah kata benda A dihubungkan dengan kata benda B dengan 'jenis dari', maka B adalah hypernym dari A atau A adalah hyponym dari B. Sebagai contoh car adalah hypernym hatchback, atau hatchback adalah hyponym dari car.

Metode kesamaan semantik perhitungan pada WordNet dibagi dalam dua kelompok besar pendekatan, yaitu *path length* dan *information content*. Beberapa contoh pendekatan dengan path length adalah *Leacock-Chodorow*, *Resnik*, *Wu-Palmer*. Beberapa contoh pendekatan dengan *Information content* adalah *Lin* dan *Jiang Conrath*. [4].

2.3 Leacock & Chodrow

Metode perhitungan kesamaan semantik pada wordnet dibagi dalam dua kelompok besar pendekatan, yaitu *path length* dan *information content.* Metode *Leacock & Chodrow* termasuk kelompok *path length*. Persamaan dari *Leacock & Chodrow* adalah sebagai berikut:

$$l\text{ch} = \log(\frac{(2 * D)}{(\text{length}(c1, c2))}) \qquad \text{H2}$$

di mana :

c1	= konsep1
c2	= konsep2
length(c1,c2)	= panjang jalur yang paling pendek (yaitu., jumlah minimum edge antara dua konsep)
D	= Maksimum depth dari taksonomi (Jumlah maksimum node dari skema ontology dari dua konsep)

Berikut ini contoh penghitungan keterkaitan antar kata *teacher-employee-worker* dengan menggunakan metode Leacok & Chodrow seperti yang dirumuskan pada persamaan F.1. Langkah-langkah yang dilakukan adalah sebagai berikut :

1. Mencari panjang jalur dari c1 dan c2 dimana c1 adalah *teacher* dan c2 adalah *employee*.
2. Masukkan c1 dan c2 ke dalam WordNet.
3. Hitung panjang jalur mulai dari entity sampai ke c1 atau c2
4. Maka kita dapatkan nilai c1 adalah 9 dan c2 adalah 7 untuk mendapatkan nilai length dari c1 dan c2, dengan mengambil nilai yang paling minimum diantara c1 dan c2 yakni 7.
5. Sedangkan untuk mencari nilai D dapat diperoleh dengan membandingkan jumlah node dari c1 dan c2, maka nilai D adalah jumlah maksimum node diantara c1 dan c2, yaitu 10.
6. Perhitungannya akan menjadi

$$\text{lch} = \log(\frac{(2 * 10)}{\text{length}(9,7)})$$

$$\text{lch} = \log(\frac{20}{7})$$

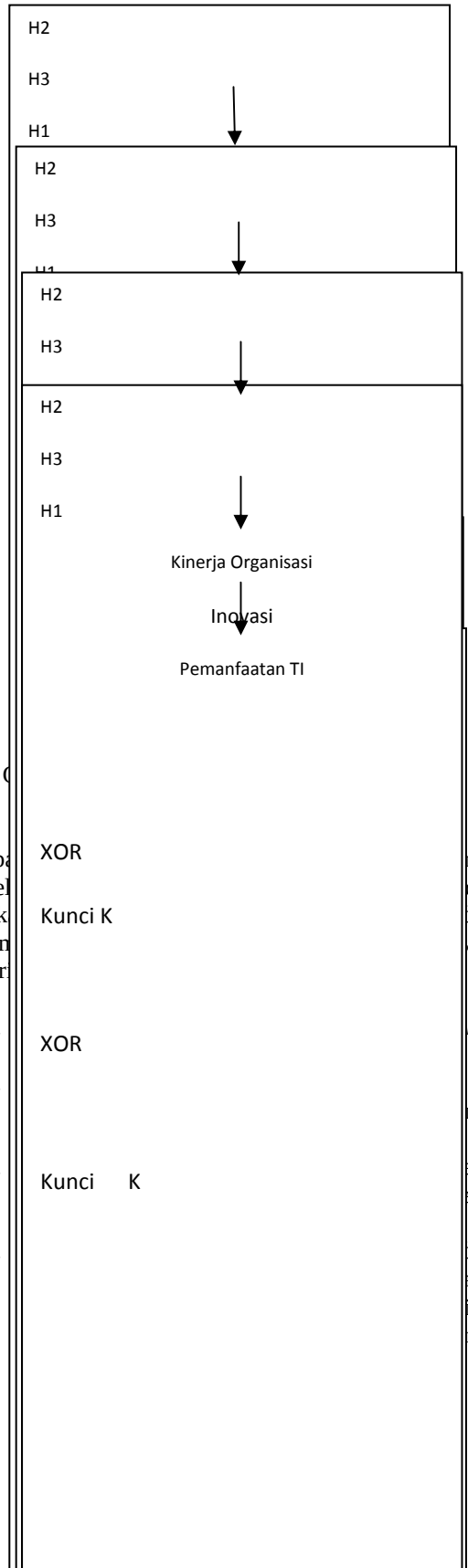
$$= 0{,}45$$

Langkah 1-6 diulang untuk mencari keterkaitan antar kata *teacher-worker* dan *employee-worker*. Sehingga didapatkan untuk *teacher-worker* adalah 0,52 dan untuk *employee-worker* = 0,90 [3].

3. Ujicoba

Penelitian pada paper ini menggunakan beberapa kata kunci sebagai contoh, diantaranya “Jumlah Mobil di Jakarta”, “Jumlah Pelajar di Jakarta”, “Jumlah Perguruan Tinggi di Jakarta”.

Gambar 3 menunjukkan langkah yang dilakukan pada pembuatan ulang query atas suatu original query.



5. Terjemahkan konsep dari bahasa Inggris ke Bahasa Indonesia, dengan tetap menyimpan besaran similaritasnya. Besaran similaritas ini nantinya berguna dalam langkah menentukan batasan similaritas yang diinginkan.Misalkan batasan similar yang diinginkan adalah 0,3 maka berdasarkan hasil perhitungan pada tabel 3 diperoleh bahwa mobil (dalam bahasa Indonesia) akan memperoleh hasil konsep vehicle, automobile dan truck.
6. Tiap-tiap konsep yang dihaslkan dalam langkah 6 diterjemahkan kembali ke bahas Indonesia, sehingga menghasilkan konsep sebagai berikut:
   - Vehicle → kendaraan
   - Automobile →mobil
   - Truck →truk
7. Pembentukan ulang query berdasarkan hasil langkah 6 adalah sebagai berikut:
   - Jumlah kendaraan di Jakarta
   - Jumlah mobil Jakarta
   - Jumlah Truck di Jakarta
   - Hitung kendaraan di Jakarta
   - Hitung mobil di Jakarta

Tabel 1. Hasil Perhitungan Menggunakan Metode Leacock & Chodrow di wordnet

Konsep 1	Konsep 2	Lch
car	vehicle	0,33
car	automobile	1
student	pupil	0,09
in	at	0,25
number	count	1
vehicle	automobile	0,2
campus	college	0,08
university	college	0,33
campus	university	0,08
academy	university	0,33
campus	academy	0,08
college	academy	0,25
truck	car	0,33
vehicle	truck	0,25
vehicle	motorcycle	0,2
motorcycle	motorbike	1
motorcycle	bicycle	0,33
bike	bicycle	1
bike	vehicle	0,33
vehicle	bicycle	0,33
plane	aircraft	0,33
vehicle	bus	0,25
student	college	0,1
campus	student	0,08
computer	notebook	0,2
laptop	notebook	0,33
laptop	computer	0,2
pc	computer	0,33
pc	notebook	0,33
laptop	Pc	0,33
pupil	campus	0,09
pupil	school	0,1
high school	pupil	0,06

Adapun ringkasan hasil ujicoba dapat dilihat pada tabel 2.

Tabel 2. Ringkasan Hasil Ujicoba

No.	Query-N	Nilai batas	Jumlah Query rewriting
1	Jumlah Mobil di Jakarta	0,2	8
2	Jumlah Mobil di Jakarta	0,3	4
3	Jumlah Mobil di Jakarta	0,8	2
4	Jumlah Pelajar di Jakarta	0,2	8
5	Jumlah Pelajar di Jakarta	0,3	4
6	Jumlah Pelajar di Jakarta	0,8	2

4. Kesimpulan

Paper ini mengajukan suatu metode baru dalam meningkatkan kinerja query rewriting dengan melibatkan pencarian sinonim konsep dan mempertimbangkan hubungan semantik diantara kandidat konsep yang akan membentuk ulang query. Hubungan semantik ditunjukkan melalui nilai similaritas yang diperoleh melalui sebuah perhitungan semantik antar konsep menggunakan tesaurus bahasa Inggris yang disebut WordNet.

Nilai similaritas digunakan sebagai pertimbangan batasan konsep yang akan dijadikan sebagai kandidat konsep untuk query rewriting. Adanya nilai similaritas inimemberikan suatu fleksibilitas terhadap luasan/cakupan kandidat konsep yang pada akhirnya juga membatasi banyaknya query rewriting yang akan dilakukan.

Penelitian ini masih merupakan penelitian awal dalam meningkatkan kinerja query rewriting dengan melibatkan pencarian sinonim konsep. Oleh karena itu perlu dilakukan penelitian lebih mendalam, diantarnya adalah bagaimana

menangani respon dari hasil query rewriting yang ada.

Daftar Pustaka:

[1] Mandala Rila, 2006, *Evaluasi Efektivitas Metode Machine Learning pada Search-Engine*, Bandung, Institut Teknologi Bandung.

[2] Richardson, Smeaton, Murphy J, 1995, *Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Words*, Dublin City University.

[3] Wicaksana I Wayan Simri, Yuniar Lintang Banowosari, Wulandari Lily, Wirawan Setia, 2006, *Pentingnya Peranan Bahasa dalam Interoperabilitas Informasi berbasiskan Komputer Karena Keragaman Semantik*, Depok, Universitas Gunadarma.

[4] Wicaksana I Wayan Simri, Hakim Reza A.., 2006, *Pendekatan Schema Matching dalam Bahasa Indonesia*, Depok, Universitas Gunadarma dan PT Radiant Centra Nusa.