QUERY EXPANSION DENGAN MENGGABUNGKAN METODE RUANG VEKTOR DAN WORDNET PADA SISTEM INFORMATION RETRIEVAL

Susetyo Adi Nugroho [(1)]

Abstrak:

Salah satu metode yang sering digunakan dalam mengukur relevansi dokumen pada sistem information retrieval adalah vector space model. Dalam pengembangan metode ini, salah satunya dapat dilakukan dengan cara melakukan perluasan terhadap vektor querynya. Perluasan dilakukan dengan menggunakan wordnet pada term-term penyusun query dengan harapan agar hasil dari sistem dapat ditingkatkan.

Kata Kunci : *information retrieval, vector space model, wordnet.*

1. Pendahuluan

Seiring dengan perkembangan informasi, disadari bahwa masalah utama telah bergeser dari cara mengakses atau bagaimana mencari informasi, namun menjadi bagaimana memilih informasi yang berguna secara selektif. Usaha untuk memilih informasi ternyata lebih besar dari sekedar mendapatkan akses terhadap informasi. Pemilihan atau penemuan kembali informasi ini tidak mungkin dilakukan secara manual karena kumpulan informasi yang sangat besar dan terus bertambah besar. Melalui penelitian dibangun sistem-sistem otomatis yang dapat membantu user dalam proses pencarian.

Penelitian ini dilakukan dengan harapan mendapatkan sebuah sistem baru yang dapat menjawab kebutuhan user, proses penelitian dilakukan dengan menggunakan vector space model. *Vector space model* adalah suatu model yang digunakan untuk mengukur kemiripan antara suatu dokumen dengan suatu query. Pada model ini, query dan dokumen dianggap sebagai vektor-vektor pada ruang n-dimensi, dimana n adalah jumlah dari seluruh term yang ada dalam leksikon. Leksikon adalah daftar semua term yang ada dalam indeks.

Salah satu cara yang dapat dilakukan untuk mengatasi hal tersebut adalah dengan menambahkan fungsi perluasan terhadap query, dimana query akan diperluas dengan menggunakan sinonim dari WordNet. Perluasan ini diharapakan dapat meningkatkan performa dari sistem , sehingga memberikan hasil yang lebih baik.

2. Wordnet

WordNet adalah suatu sistem referensi leksikal bahasa inggris yang bersifat online. WordNet dikembangkan oleh Cognitive Science Laboratory di Universitas Princeton yang dikepalai oleh George Miller. Arti dari suatu kata pada WordNet direpresentasikan dengan synonym sets (synsets). Synsets adalah daftar *term* atau *collocation* yang artinya sama dan dalam konteks tertentu penggunaannya dapat saling dipertukarkan. Dalam synset juga dicatat pointer-pointer ke synset lain yang digunakan untuk mendeskripsikan relasi antar synset. WordNet dibagi dalam empat taksonomi berdasarkan type kata yaitu kata benda, kata kerja, kata keterangan, dan kata sifat (Miller, 1990).

3. Query Expansion

Query Expansion atau perluasan query adalah proses me-reformulasikan kembali query awal dengan melakukan penambahan beberapa *term* atau kata pada *query* untuk meningkatkan perfoma dalam proses information retrieval. Dalam konteks web *search engine*, hal ini termasuk evaluasi input user dan memperluas query pencarian untuk mendapatkan dokumen yang cocok dengan query (Qiu, 1993). Proses perluasan dalam sistem ini dilakukan dengan menggunakan sinonim dari wordnet. Metode yang dilakukan dalam perluasan adalah dengan mencari sinonim dalam bentuk *unstemmed-term* dari query. Pencarian sinonim tidak memperhatikan tiap relasi

[(1)] *Susetyo Adi Nugroho, Mahasiswa* Program Studi Teknik Informatika Fakiltas Teknik Universitas Kristen Duta Wacana.

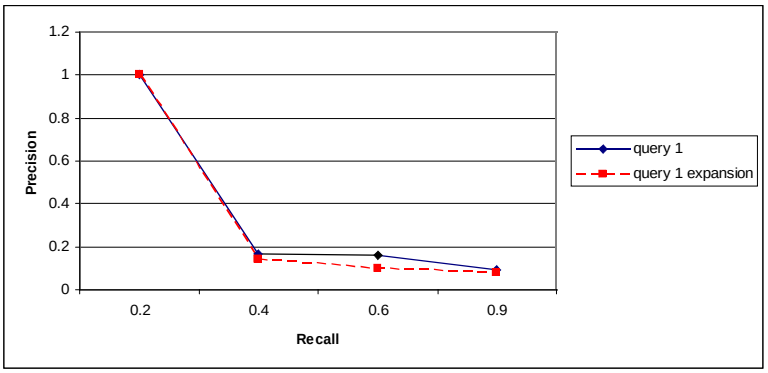
dari synset yang ditemukan dalam wordnet, dan hanya akan diambil maksimal 5 sense dari tiap term yang sinonimnya ditemukan.

4. Pengujian

Proses uji coba dilakukan dengan menggunakan koleksi data test yang sering digunakan dalam proses uji coba sistem IR, yaitu ADI (American Documentation Institute) test collection. Seluruh koleksi dari dokumen dan query dalam bahasa inggris. Proses *indexing* 82 koleksi dokumen memakan waktu kurang lebih 4-5 menit. Pengujiannya dilakukan dengan 6 buah query dengan panjang query yang berbeda.

4.1 Pengujian 1

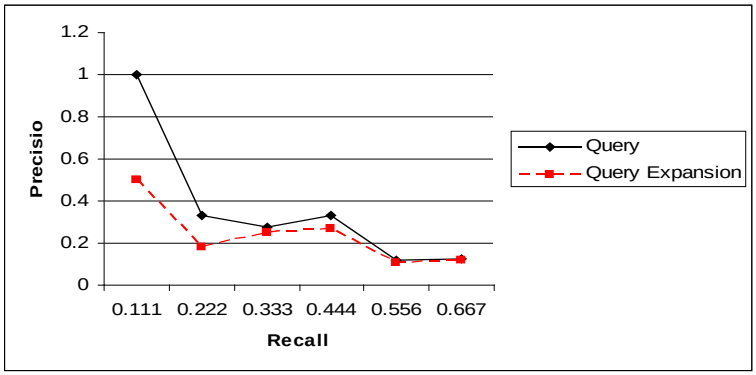
Query untuk pengujian 1 adalah “the use of abstract mathematics in information retrieval, e.g. group theory”. Hasil yang relevan dari query ini berjumlah 5 dokumen. Proses perluasan query 1 merubah hasil karena adanya perubahan rangking akibat recall naik. Nilai precission relatif turun pada query 1.



Gambar 1 Grafik precision dokumen terhadap query 1

4.2 Pengujian 2

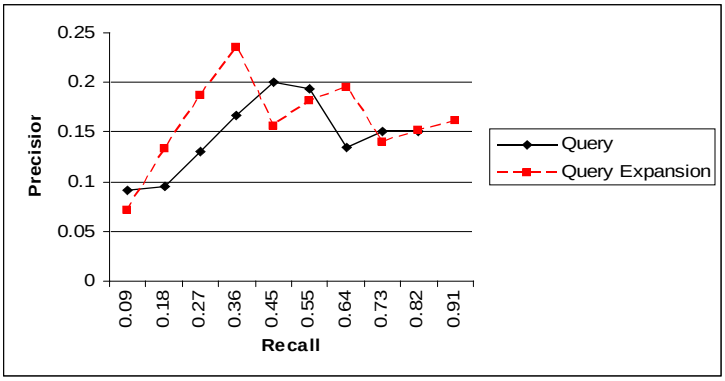
Pengujian kedua di lakukan dengan query “information dissemination by journals and periodicals”. Query ini akan mengembalikan 6 dari 9 dokumen relevan. Perluasan terhadap query ini justru memperburuk nilai precission terhadap dokumen yang terjadi karena adanya perubahan ranking.



Gambar 2 Grafik precision dokumen terhadap query 2

4.3 Pengujian 3

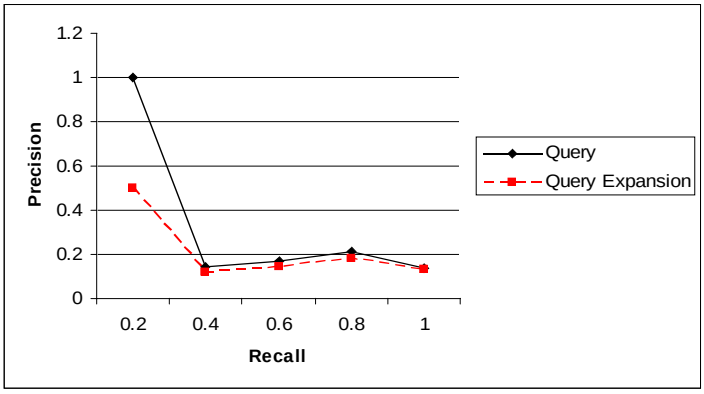
Query untuk pengujian 3 adalah “Information systems in the physical sciences”. Hasil yang relevan dari query ini berjumlah 11 dokumen. Proses perluasan pada query 3 memberikan pengaruh positif, disebabkan rangking dokumen relevan naik.



Gambar 3 Grafik precision dokumen terhadap query 3

4.4 Pengujian 4

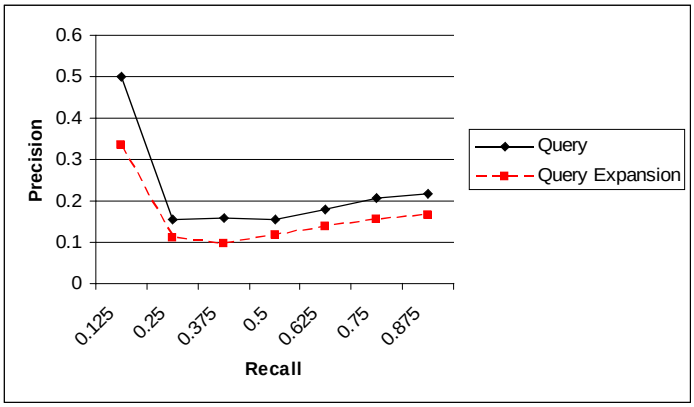
Pengujian keempat di lakukan dengan query "Methods of coding used in computerized index systems". Proses pencarian akan menghasilkan 5 buah dokumen relevan. Query expansion pada query 4 tidak memberikan hasil yang lebih baik, ini terjadi karena hampir seluruh term perluasan tidak ada dalam index.



Gambar 4. Grafik precision dokumen terhadap query 4

4.5 Pengujian 5

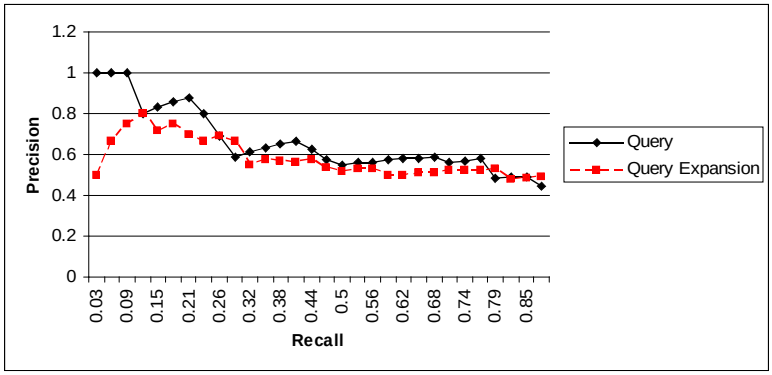
Pengujian kelima di lakukan dengan query "Government supported agencies and projects dealing with information dissemination". Proses pencarian dengan query ini akan mengembalikan 7 dari 8 buah dokumen relevan. Adanya kenaikan recall pada perluasan query 5 membuat ranking dokumen relevan turun, akibatnya precission dokumen hasil perluasan query lebih kecil dari query awal.



Gambar 5. Grafik precision dokumen terhadap query 5

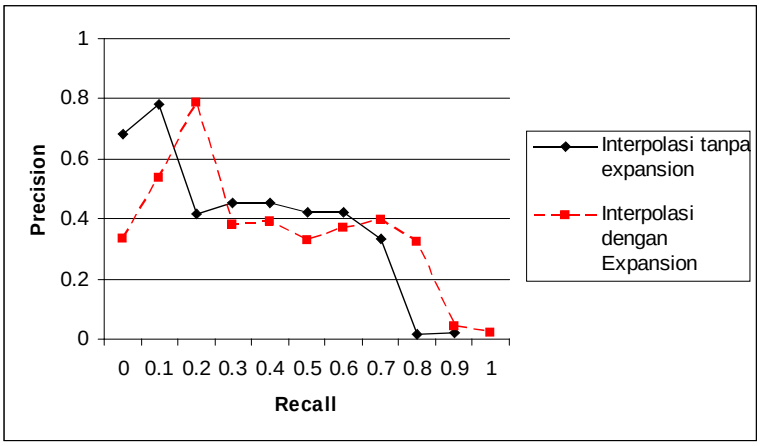
4.6 Pengujian 6

Query untuk pengujian 6 adalah "computerized information retrieval systems. computerized indexing systems". Hasil yang relevan dari query ini berjumlah 34 dokumen. Hasil perluasan meningkat hanya pada level 0.27 dan 0.8, ini dikarenakan pada level recall tersebut, rangking dokumen naik dari sebelumnya.



Gambar 6 Grafik precision dokumen terhadap query 6

Gambar 7 menunjukkan grafik rata-rata interpolasi antara 2 proses query untuk semua query. Pada beberapa level recall, nilai precission lebih tinggi karena pada level tersebut jumlah dokumen relevan lebih banyak terambil oleh sistem.



Gambar 7 Grafik Interpolasi Recall /precision terhadap seluruh query

5. Kesimpulan

Dari data dan hasil pengujian query terhadap sistem, baik tanpa maupun dengan query expansion, dapat disimpulkan hasil penelitian yang dilakukan dengan melakukan query expansion menggunakan sinonim dari wordnet pada metode ruang vektor adalah sebagai berikut:

- Penggunaan query expansion berhasil meningkatkan jumlah dokumen yang diterima oleh sistem.
- Sistem dengan perluasan query tidak menaikkan nilai precision karena rangking dokumen relevan yang dikembalikan turun. Rangking turun karena semakin banyak dokumen non-relevan yang diterima oleh sistem.
- Penggunaan sinonim dari WordNet untuk memperluas query dengan mengambil part of speech noun bagian sinonim tidak membantu dalam meningkatkan nilai precision. Hal ini terjadi karena metode pengambilan sinonim tiap query tanpa memperhitungkan keterkaitan relasi dan derajat kesamaan dengan term query yang dimaksud.

6. Saran

- Menambahkan kemampuan untuk mengenali dan menggunakan keterkaitan relasi dan derajat kesamaan antar sysnset dari wordnet dalam proses perluasan query.
- Menggunakan file wordnet selain bagian sinonim kategori noun. Hal ini disebabkan banyak sinonim kata yang sering muncul pada kategori lain seperti hypernim atau hiponim, yang mungkin lebih cocok dengan term yang dimaksud.
- Menambahkan kemampuan melakukan pembetulan penulisan terhadap query jika terjadi kesalahan penulisan (spelling errors).

7. Daftar Pustaka

Buscaldi, Davide dan Paolo Rosso dan Emilio Sanchis Arnal. 2005. A WordNet-based Query Expansion method for Geographical Information Retrieval. Universidad Polit´ecnica de Valencia, Spain.

Grosman , David A dan Frieder O.2004. *Information Retrieval: Algorithm and heuristics, 2nd Edition*. Springer.

Haryono , M.E.H. 2005. Query expansion menggunakan model perpaduan genetika dan hand-crafted thesaurus. *SNIKTI VI 2005*.E7-E11.

Mandala dan Setiawan. 2004. Improving Information Retrieval System Performance by Automatic Query Expansion. *Jurnal ITB*, Bandung.

Mandala, Rila, dan Tokunanga Takenobu, dan Tanaka Hozumi. 1998. The Use of WordNet in Information Retrieval. Department of Computer Science Tokyo Institute of Technology.

Miller, G.A. 1990. Introduction to WordNet: An On-line Lexical Database. International Journal of Lexicography, Vol. 3, pp. 235-312.

Ribero-Neto, Berthier dan Ricardo Baeza-Yates. 1999. Modern Information Retrieval. ACM Press: New York.

Qiu, Y. And Rfe, HP.1993. Concept-based query expansion. *SIGIR '93*, hal 160-169.

Voorhees, Ellen M. 1993. Using wordnet to disambiguate word sense for text retrieval. Procedings of the 16th ACM-SIGIR Conference, hal. 171-180.

Voorhees, Ellen M dan Yuan-Wang Hou. Vector Expansion in a Large Collection. Siemens Corporate Research, Inc. http://trec.nist.gov/pubs/trec1/papers/27
.txt, tanggal akses 7 Juli 2008.