PENCARIAN DOKUMEN BERDASARKAN KOMBINASI ANTARA MODEL RUANG VEKTOR DAN MODEL DOMAIN ONTOLOGI

Agung Hadhiatma
Jurusan Teknik Informatika, Fakultas Sains dan Teknologi,
Universitas Sanata Dharma Yogyakarta
Paingan, Maguwoharjo, Depok, Sleman
e-mail: agunghad@yahoo.com

Abstrak

Selama ini pencarian dokumen (information retrieval) pada mesin pencari adalah berdasarkan kata kunci (keyword) sebagai input untuk proses pencarian. Ada banyak metode yang digunakan untuk proses pencarian ini, antara lain: term weighting, boolean model dan vector model. Hasil dari pencarian dokumen dengan menggunakan metode ini kadang belum tentu sesuai yang diinginkan, meskipun dalam dokumen yang diperoleh tersebut sudah mengandung kata kunci dalam masukan keyword. Salah satu penyebabnya adalah adanya berbagai konsep domain pengetahuan yang berbeda dalam kata kunci yang sama (problem semantik).

Penelitian ini bertujuan meningkatkan ketepatan mesin pencari dalam pencarian dokumen agar hasil pencarian dokumen lebih mendekati dengan kebutuhan pengguna. Studi kasus yang digunakan adalah untuk sumber pustaka digital pewayangan.

Pencarian dengan input kata kunci menggunakan model ruang vektor dan pencarian secara semantik menggunakan model domain ontologi. Pencarian dengan input kata kunci menggunakan indeks dokumen berdasarkan pembobotan/term weighting sedangkan pencarian secara semantik menggunakan indeks dokumen berdasarkan metadata domain ontologi. Domain ontologi tersebut merupakan ekstraksi pengetahuan dari dokumen. Pada pencarian semantik akan dimunculkan pilihan kombinasi kata kunci pada konsep domain yang berbeda. Kombinasi kata kunci tersebut merupakan perluasan pada kata kunci yang dimasukkan oleh pengguna.

Kata kunci*: pencarian dokumen, model ruang vektor, model domain ontologi, perluasan kata kunci*

1. PENDAHULUAN

Selama ini pencarian dokumen (*information retrieval*) pada mesin pencari adalah berdasarkan kata kunci (keyword) sebagai input untuk proses pencarian. Ada banyak metode yang digunakan untuk proses pencarian ini, antara lain: term weighting, boolean model dan vector model. Hasil dari pencarian dokumen dengan menggunakan metode ini kadang belum tentu sesuai yang diinginkan, meskipun dalam dokumen yang diperoleh tersebut sudah mengandung kata kunci dalam masukan keyword. Salah satu penyebabnya adalah adanya berbagai konsep domain pengetahuan yang berbeda dalam kata kunci yang sama (problem semantik). Sebagai contoh adalah pencarian dengan kata kunci "banteng". Hasil pencarian dari kata kunci tersebut dapat merupakan dokumen mengenai hewan banteng ataupun dokumen mengenai berita salah satu partai tertentu yang berlambang banteng. Sebuah kata yang sama mengandung konteks atau domain pembicaraan yang berbeda. Pencarian dengan kata kunci, masih menuntut kita untuk memilih dokumen yang relevan dengan kebutuhan kita. Meskipun hasil pencarian dokumen sudah diranking berdasarkan kedekatan dengan kata kunci yang diberikan, ranking dokumen yang diperoleh belum tentu sesuai dengan yang diharapkan oleh pengguna. Untuk itu pencarian dokumen membutuhkan sebuah kedalaman pengetahuan dari pengguna untuk menjelaskan atau menambahkan keterangan pada sebuah kata.

Penelitian ini bertujuan meningkatkan ketepatan mesin pencari dalam pencarian dokumen agar hasil pencarian dokumen lebih mendekati dengan kebutuhan pengguna. Mesin pencari membantu pengguna dengan memberikan alternatif beberapa pilihan perluasan kata kunci. Masing-masing perluasan kata kunci merupakan representasi dari suatu domain pengetahuan/konsep/konteks. Studi kasus yang digunakan adalah untuk sumber dokumen digital cerita wayang.

2. TINJAUAN PUSTAKA

2.1. Pencarian Dokumen

Sistem temu kembali informasi (*information retrieval system*) digunakan untuk menemukan kembali (*retrieve*) secara otomatis informasi- informasi yang relevan terhadap kebutuhan pengguna dari suatu kumpulan informasi. Banyak peneliti telah melakukan penelitian dalam pencarian dokumen. Sekarang ini penelitian dilakukan untuk meningkatkan ketepatan pencarian, salah satunya adalah melakukan penelitian berdasarkan konteks (Ozcan dan Aslandogan, 2003). Beberapa model yang telah dibuat adalah dengan menggunakan *similarity* antara konsep ontologi domain dan ontologi *query* (Araujo dan Pinto, 2007). Beberapa dari model tersebut menggunakan teori graph dan metode formal yang kompleks (Thiagrajaan dkk, 2009), (Guarino, 1998). Penelitian ini dilakukan dengan mengajukan model yang lebih sederhana dalam hal pencarian berdasarkan konteks atau domain pembicaraan pada dokumen cerita dengan fokus pada cerita wayang. Model

yang diajukan adalah penggabungkan teknik klasik *information retrieval* yaitu model ruang vektor dengan perluasan *query* pada model ontologi.

2.2 Model Ruang Vektor

2.2.1. Pengindeksan

Sistem temu kembali informasi terbagi dalam 2 proses, yaitu proses indexing dan proses perankingan dokumen. Adapun tahapan dari proses pengindeksan adalah sebagai berikut :

- *Parsing* dokumen yaitu proses pengambilan kata-kata dari kumpulan dokumen.
- *Stoplist* yaitu proses pembuangan kata buang seperti: tetapi, yaitu, sedangkan, dan sebagainya.
- *Stemming* yaitu proses penghilangan/ pemotongan dari suatu kata menjadi bentuk dasar. Kata "diadaptasikan" atau "beradaptasi" mejadi kata "adaptasi" sebagai istilah.
- Menghitung term frekuensi (TF) dan dokumen frekuensi (IDF), kemudian dimasukkan ke *database* indeks
-

2.2.2. Pembobotan TF-IDF

Kata dalam dokumen diberi bobot, Pembobotan tersebut berdasarkan pada rumus Pembobotan TF-IDF sebagai berikut (Maning, dkk, 2008):

$W_{ij} = TF_{ij} * IDFj$, dimana $IDFi = \log (n/DF_i)$, dimana :

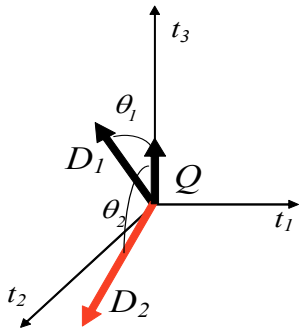
- W_{ij} = bobot istilah kata i pada dokumen j.
- TF_{ij} = frekuensi istilah kata i dalam dokumen j.
- n = jumlah dokumen.
- DF_i = jumlah dokumen yang mengandung istilah kata i.

2.2.3. Similarity pada Ruang Vektor

Koleksi dokumen direpresentasikan dalam ruang vektor sebagai matriks kata-dokumen (*terms-documents matrix*). Nilai dari elemen matriks w_{ij} adalah bobot kata *i* dalam dokumen *j*. Misalkan terdapat sekumpulan kata *T* sejumlah *n*, yaitu $T = (T_1, T_2, \ldots, T_n)$ dan sekumpulan dokumen *D* sejumlah *m*, yaitu $D = (D_1, D_2, \ldots, D_m)$ serta w_{ij} adalah bobot kata *i* pada dokumen *j*. Maka representasi matriks kata-dokumen adalah :

$$\begin{array}{c} \\ D_1 \\ D_2 \\ \vdots \\ \vdots \\ D_m \end{array} \begin{array}{c} \begin{array}{cccc} T_1 & T_2 & .... & T_n \end{array} \\ \begin{bmatrix} w_{11} & w_{21} & \ldots & w_{n1} \\ w_{12} & w_{22} & \ldots & w_{n2} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ w_{1m} & w_{2m} & \ldots & w_{nm} \end{bmatrix} \end{array}$$

Penentuan relevansi dokumen dengan *query* dipandang sebagai pengukuran kesamaan (*similarity measure*) antara vektor dokumen (D) dengan vektor *query* (Q). Semakin "sama" suatu vektor dokumen dengan vektor *query* maka dokumen dapat dipandang semakin relevan dengan *query*. Ukuran kesamaan ditunjukan dengan besarnya sudut θ. Semakin kecil sudut berarti dapat dianggap semakin mirip (relevan). Dokumen D_1 lebih mirip dengan *query* Q daripada dokumen D_2



Gambar 2.1 Sudut yang dibentuk antara dokumen dan vektor pada ruang vektor

Jika *Q* adalah vektor *query* dan *D* adalah vektor dokumen, yang merupakan dua buah vektor dalam ruang berdimensi-*n*, dan *θ* adalah sudut yang dibentuk oleh kedua vektor tersebut. Maka:

Inner product : $Q \bullet D = |Q||D| \cos\theta$ $\quad |D| = \sqrt{\sum_{i=1}^{n} Di^2}$ dan $|Q| = \sqrt{\sum_{i=1}^{n} Qi^2}$

Rumus yang digunakan untuk memgukur jarak kedekatan antar vektor adalah sebagai berikut :

$$Sim(Q,D) = \cos(Q,D) = \frac{Q \bullet D}{|Q||D|} = \frac{1}{|Q||D|} \sum_{i=1}^{n} Qi \bullet Di$$

Kedekatan *query* dan dokumen diindikasikan dengan sudut yang dibentuk. Nilai cosinus yang cenderung besar mengindikasikan bahwa dokumen cenderung sesuai *query*. Nilai cosinus sama dengan 1 mengindikasikan bahwa dokumen sesuai dengan dengan *query*.

2.3. Ontologi

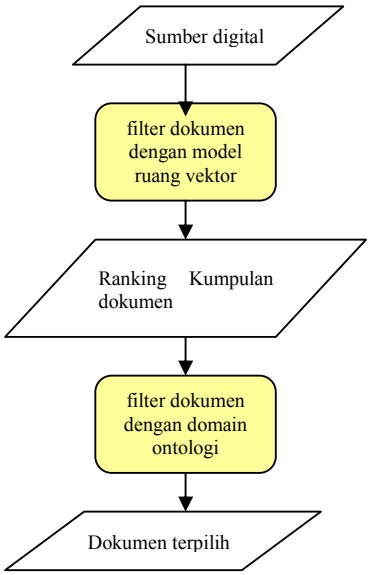
Ontologi merupakan sebuah spesifikasi eksplisit dari sebuah konseptualisasi (Gruber, 1995). Dengan kata lain, ontologi adalah sebuah konsep yang secara sistematik menjelaskan segala sesuatu yang ada. Menurut Finin (1991) ontologi merupakan definisi dari pengertian dasar dan relasi kosakata dari sebuah domain dengan menggunakan aturan dari kombinasi istilah dan relasi. Dalam bidang *Artificial Intelligence* (AI) ontologi memiliki dua pengertian yang berkaitan. Pertama ontologi merupakan kosakata representasi yang sering dikhususkan untuk domain atau subjek pembahasan tertentu. Kedua sebagai suatu *body of knowledge* untuk menjelaskan suatu bahasan tertentu. Menurut Gruber Sebuah ontologi dijelaskan dengan menggunakan notasi dari konsep (kelas), *instances*, relasi, fungsi, dan aksiom. Sedangkan Borst melakukan modifikasi dari definisi Gruber dengan mengatakan "Sebuah ontologi adalah spesifikasi formal dari sebuah konseptual yang diterima (*shared*)". Sebuah badan yang merupakan konsorsium world wide web membuat sebuah bahasa ontologi yang dapat dipakai untuk representasi pengetahuan. Bahasa ontologi tersebut adalah OWL (Ontology Web Language) OWL ini merupakan pengembangan dari RDF (Resource Description F$ramework).

- RDF (Resource Description Framework)
  RDF merupakan pengembangan dalam metadata XML. Tidak seperti pada XML, di dalam RDF tidak sekedar direpresentasikan sebuah struktur metadata tetapi juga direpesentasikan struktur relasi antara entitas-entitas. Relasi tersebut dapat membentuk juga unsur subyek, predikat dan obyek. RDF menunjukkan atau mempresentasikan realitas dengan memformalkan model grafis dengan XML sintaks + semantik. RDF Scema (RDFS) adalah pengembangan RDF dengan skema kosakata dalam bentuk: class, propery type, subclassOf, subPropertyOfRange, domain.
- OWL (Ontology Web Language)
  Pengembangan lebih lanjut dari RDF yang ditempuh dengan menambahkan beberapa kosakata untuk menjelaskan *properties* dan *classes*. Antara lain terwujud dalam relasi antara *classes*, kardinalitas, *equality*, berbagai tipe dari *properties*, karakteristik dari *properties*.

RDF sebagai representasi pengetahuan dalam ontologi tersebut dapat dikueri (*querying*) dengan bahasa kueri RDQL dengan bahasa java (package jena).

3. METODE PENELITIAN

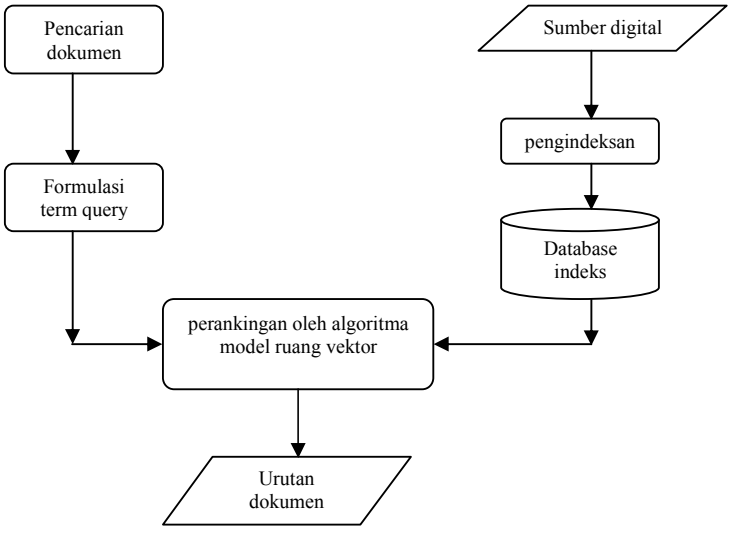
Bagian penting dari penyelesaian masalah pencarian dokumen adalah: bagaimana dokumen- dokumen dalam sumber digital dibuat indeksnya dan bagaimana proses pencarian tersebut dilakukan. Untuk itu diusulkan Proses filter/penyaringan dokumen dalam dua tingkat. (gambar 3.1).



Gambar 3.1: Proses filter dokumen 2 tingkat.

3.1. Filter dokumen dengan model ruang vektor

Gambar 3.2 memperlihatkan bahwa terdapat tiga langkah operasi pada sistem temu kembali informasi. Langkah pertama dimulai dari koleksi dokumen dalam bentuk sumber digital (dapat dilihat dalam panah) sampai pada proses terbentuknya *database* indeks. Langkah kedua dimulai dari *query* pencarian dokumen oleh pengguna. Dalam query tersebut akan dilakukan formulasi term query, yaitu penghitungan bobot dari term-term query tersebut dengan mengunakan algoritma pembobotan TF- IDF. Sedangkan langkah ketiga adalah proses perankingan dokumen dengan menggunakakan algoritma model ruang vektor.

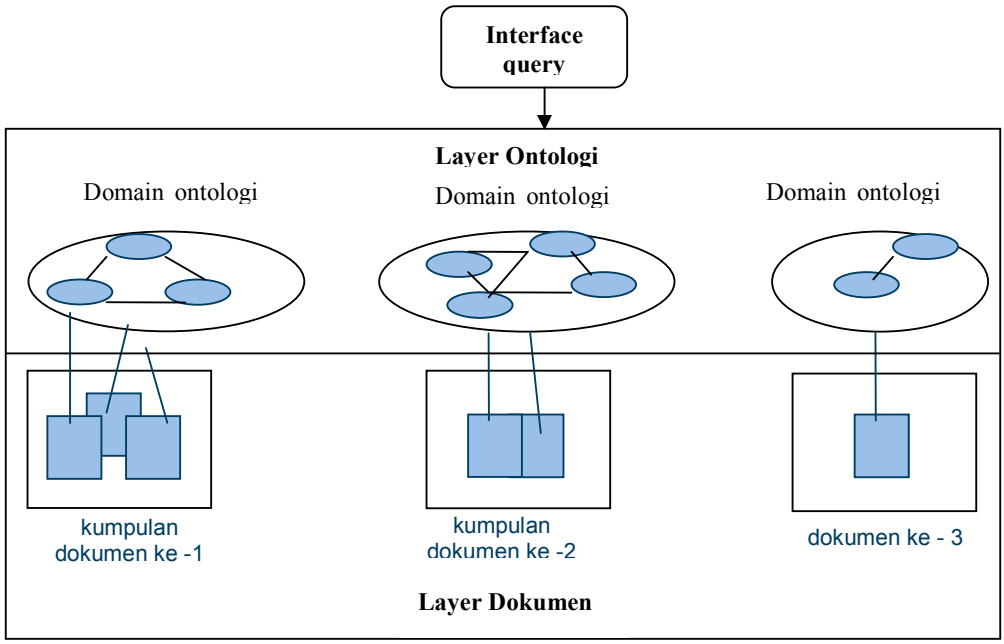


Gambar 3.2: filter dokumen dengan model ruang vektor.

Dalam model ruang vektor, *query* dan dokumen mempunyai vektor bobot untuk kata (term). Kesamaan dokumen dan *query* dihitung berdasarkan jarak vektor dokumen (D) dan query (Q) dengan menggunakan rumus *inner product* (*dot product*). Jarak tersebut dihitung dengan ukuran sudut cosinus.

3.2. Filter dokumen dengan domain ontologi

Secara arsitektur diusulkan dalam dua layer: layer dokumen dan layer domain ontologi. Layer dokumen merupakan *database* dokumen, sedangkan layer domain ontologi merupakan sebuah penjelasan/catatan/ abstraksi/ekstraksi metadata dari layer dokumen. Ekstraksi atau penulisan abstraksi dokumen dalam bentuk /representasi domain ontologi tersebut dilakukan secara manual dengan menggunakan editor untuk pembuatan ontologi yaitu: protégé. Representasi ontologi menggunakan RDF (Resource Description Framework) dan *query* menggunakan RDQL (bahasa *query* untuk RDF)



Gambar 3.3: filter dokumen dengan domain ontologi.

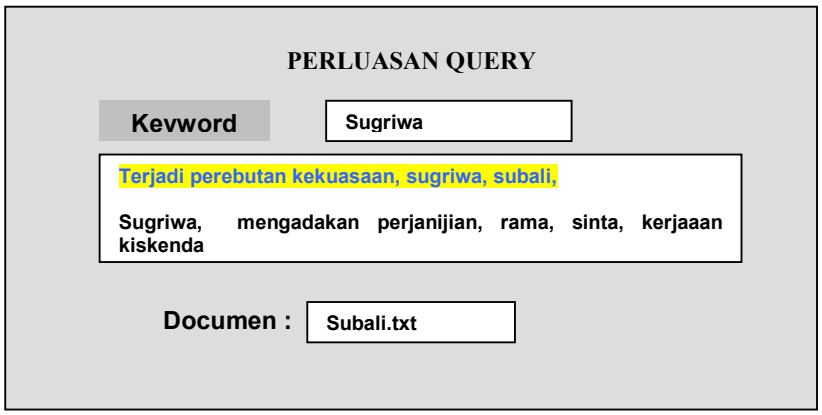
4. HASIL DAN PEMBAHASAN

Berikut ini akan disajikan hasil dari pencarian dua tahap. Untuk Tahap pertama adalah hasil dari pencarian dengan model pembobotan TF-IDF dan pencarian model ruang vektor. Input adalah kata kunci yang berhubungan dengan wayang. Kata kunci yang diinputkan adalah subali dan sugriwa. Hasil adalah sebagai berikut (gambar 4.1).

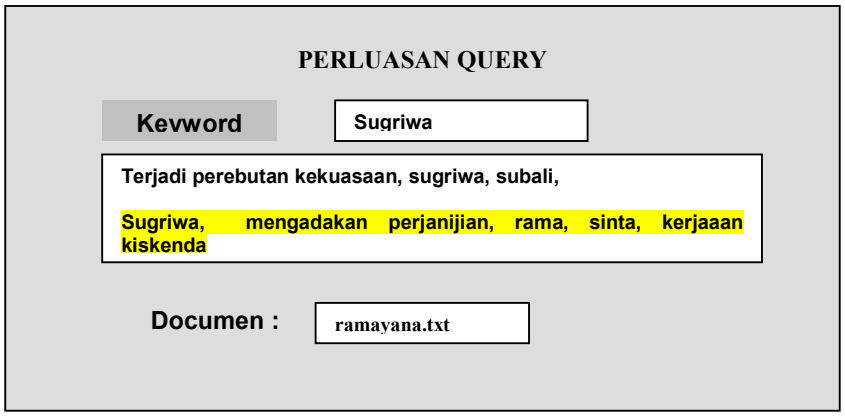
Gambar 4.1. Hasil pencarian dengan kata kunci subali dan sugriwa.

Hasil dari pencarian tersebut memunculkan beberapa dokumen berdasarkan perangkingan, antara lain: subali.txt, sugriwa.txt, rama.txt, ramayana.txt, hanoman.txt, anggodo.txt, kumbakarna.txt, rahwana.txt.
Beberapa dokumen tersebut memunculkan beberapa kisah-kisah yang berbeda atau konteks crita yang berbeda dari si tokoh subali. Pengguna awam kadang hanya mempunyai pengetahuan yang sedikit mengenai tokoh – tokoh wayang, dan kadang pula tidak ingat akan istilah dan tokoh wayang dalam berbagai kisah pewayangan. Untuk memilih salah satu dokumen yang dikehendaki pengguna berdasarkan suatu konteks cerita tertentu yang berhubungan dengan subali akan membutuhkan waktu yang lama, karena pengguna harus membaca semua dokumen yang muncul dan setelah membaca semua dokumen tersebut pengguna dapat memilih dokumen yang sesuai.

Pada pencarian tahap kedua, dokumen hasil dari pencarian tahap pertama akan digunakan sebagai masukan untuk query dalam berbagai domain pengetahuan mengenai subali. *Query* tersebut akan menampilkan term-term istilah dalam tiap domain pengetahuan sebagai perluasan kata kunci. Dari berbagi term-term dalam berbagi domain pengetahuan tersebut, pengguna dapat memilih salah satu yang sesuai dengan kebutuhannya. Hasil pencarian tahap kedua dapat dilihat pada gambar 4.2 dan gambar 4.3 . Pada gambar terlihat untuk keyword kata kunci sugriwa akan muncul perluasan kata kunci: "Terjadi perebutan kekuasaan, sugriwa, subali" dan perluasaan kata kunci: "sugriwa, mengadakan perjanjian, rama, sinta, kerjaan kiskenda". Hasil pilihan pada gambar 4.2 memunculkan hasil: subali.txt. Hasil pilihan pada gambar 4.3 memunculkan hasil ramayana.txt.

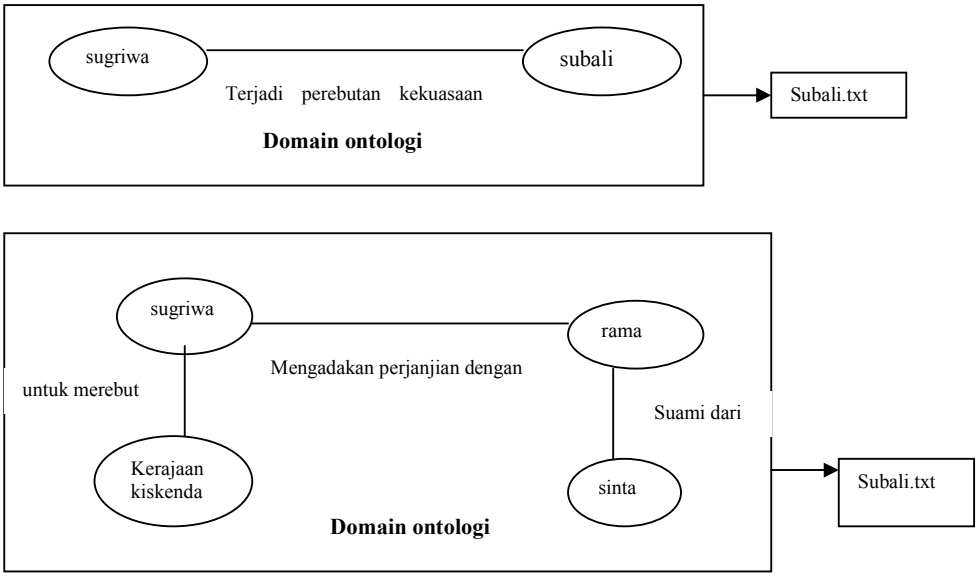


Gambar 4.3: Perluasan query dengan hasil subali.txt



Gambar 4.3: Perluasan query dengan hasil ramayana.txt

Struktur ontologinya adalah sebagai berikut: gambar 4.4



Gambar 4.4: Struktur domain ontologi

Dalam penelitian ini masih menyisakan problem sebagai berikut: pembuatan domain ontologi masih dilakukan secara manual. Pembuatan domain dilakukan secara manual dengan mengekstrak intisari dari dokumen-dokumen wayang. Ekstraksi seperti ini akan membutuhkan waktu yang lama dan membutuhkan orang yang sudah paham akan cerita wayang. Untuk penelitian kedepannya ekstraksi intisari dari dokumen dapat dilakukan secara otomatis dengan *mechine learning* melalui proses *knowledge discovery.* Ada berbagai skenario proses pembelajaran yang dapat dikembangkan untuk pembentukan ontologi secara otomatis [Davies, 2006]

5. KESIMPULAN

Kombinasi pencarian model ruang vektor dengan model domain ontologi dapat mendapatkan dokumen sesuai dengan domain pembicaraan / konteks yang diharapkan oleh pengguna. Supaya pencarian dokumen dapat sesuai dengan konteks pembicaraan yang diinginkan oleh pengguna, maka dilakukan filter dokumen dalam dua tahap. Tahap pertama dilakukan dengan input kata kunci. Pada tahap kedua: hasil dari dokumen tersebut akan memunculkan beberapa domain pembicaraan/konteks dalam bentuk perluasan *query* kata kunci. Pengguna dapat memilih kata kunci yang lebih mendekatkan dengan harapan dari pengguna. Hal ini untuk membantu pengguna dalam memilih kombinasi atau perluasan kata kunci yang lebih tepat. Domain ontologi yang dibentuk masih dilakukan secara manual. Untuk penelitian kedepannya dapat dikembangkan pembentukan domain ontologi secara otomatis dengan mengekstraksi domain pengetahuan dari dokumen tersebut dengan *knowledge discovery* ataupun *text mining.*

6. DAFTAR PUSTAKA

Davies, John., et al, 2006, *Semantic Web Technologies Trends and Research in Ontology-based Systems*, John Wiley & Sons Ltd, Chichester, West Sussex, PO19 8SQ, England.

Finin, T., *et al*. 1991, *Enabling Technology for Knowledge Sharing*, AI Magazine.

Gruber, T. R. 1995, *Towards Principles for the Design of Ontologies Used for Knowledge Sharing*, International Journal of Human-Computer Studies, 43:907-928.

Guarino , Nicola, 1998, *Formal Ontology and Information Systems*, Formal Ontology in Information Systems Proceedings of FOIS'98, Trento, Italy, Amsterdam, IOS Press, pp. 3-15.

Ozcan, Rifat, dan Aslandogan, Y. Alp, 2004, *Concept Based Information Access Using Ontologies and Latent Semantic Analysis* **,** Department of Computer Science and Engineering University of Texas , Arlington

Thiagarajan, Rajesh., et al, *Computing Semantic Similarity Using Ontologies*, 2008, the International Semantic Web Conference (ISWC), 2008, Karlsruhe, Germany.