Jurnal Teknologi, 46(D) Jun 2007: 103–120

PENDEKATAN ONTOLOGI DALAM CAPAIAN DAN PERWAKILAN SEMANTIK DOKUMEN WEB

ARIFAH CHE ALHADI[1], SHAHRUL AZMAN NOAH[2] &
LAILATUL QADRI ZAKARIA[3]

Abstrak. Visi web semantik membolehkan capaian maklumat dilakukan secara semantik, yang mana model semantik kueri dipadankan dengan maklumat semantik dokumen web. Namun demikian kebanyakan dokumen web adalah tidak berstruktur dan tidak mempunyai maklumat semantik dokumen menyebabkan kesukaran proses pengkuerian. Oleh itu, capaian dan pengekstrakan maklumat semantik daripada dokumen web adalah amat penting dalam merealisasikan visi web semantik dan meningkatkan kualiti capaian maklumat. Kertas kerja ini membincangkan pengaplikasian pendekatan ontologi spesifik dan pemprosesan bahasa tabii dalam menyokong capaian dan pengekstrakan maklumat semantik dokumen web. Dengan menggunakan kedua-dua teknik ini, setiap kali capaian maklumat dilakukan, sistem akan menjana model integrasi semantik dokumen iaitu dokumen yang dicapai oleh enjin gelintar komersial yang ditetapkan. Model intergrasi semantik dokumen ini membolehkan pengguna mencapai dan melayarinya secara semantik. Hasil pengujian capaian dan padanan konsep yang dijalankan memperlihatkan kedua-dua teknik yang digunakan mampu mengenal pasti dan mengekstrak maklumat semantik daripada kueri dan kandungan dokumen web.

Kata kunci: Capaian dokumen semantik, web semantik, ontologi, analisis bahasa tabii, perwakilan semantik dokumen, perwakilan semantik kueri

Abstract. The Semantic Web vision offers the potential to express queries in a more semantically way whereby semantic query model will be matched with semantic information of the document. However, the unstructured natures of existing web documents, which lack of semantic prove to be a difficult task for such a query. Therefore, semantic information retrieval and semantic information extraction of web documents content are important to realize semantic web vision and enhance the quality of information access. To support this, the semantic information content of web documents need to be specified in order to make the tangled information more structured and accessible. In this paper, we propose an approach meant to semantically query web documents using natural language analysis technique and a domain specific ontology. Using both techniques, the tool gradually constructs the semantic document integration model of the documents retrieved from an existing search engine for each search session. The semantic model can then be semantically refined and browsed by the user. The result of concept matching and accessing shows that both techniques that have been used could identify and extract semantic information from query and web document content.

Keywords: Semantic document retrieval, semantic web, ontology, natural language analysis, semantic document representation, semantic query representation

[1] Jabatan Sains Komputer, Fakulti Sains dan Teknologi, Universiti Malaysia Terengganu (UMT), 21030 Kuala Terengganu, Terengganu, Malaysia
Tel: 09-6683352, Faks: 09-6694660, Email: arifah_hadi@kustem.edu.my.

[2&3] Jabatan Sains Komputer, Jabatan Sains Maklumat, Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 UKM, Bangi, Selangor, Malaysia
Tel: 03-89216182, Faks: 03-89256732, Email: {samn, laila}@ftsm.ukm.my.

1.0 PENGENALAN

Kekurangan maklumat semantik pada dokumen HTML menyukarkan usaha pembangunan sistem capaian dokumen web secara semantik. *Markup* dokumen HTML hanya menyediakan struktur maklumat dokumen bukannya maklumat semantik dokumen. Namun begitu, kebanyakan kandungan web pada masa ini mengekodkan maklumat dalam bentuk yang difahami oleh manusia dan bukannya mesin. Mesin memerlukan maklumat tambahan untuk menghubungkan kandungan (semantik) dengan data [1].

Terdapat dua pendekatan yang dicadangkan untuk menambah nilai semantik dokumen [2] iaitu dengan memperkemaskan sumber maklumat dengan menotasikan maklumat semantik dalam bentuk yang boleh dicapai oleh mesin atau menulis suatu aturcara untuk mengekstrak maklumat semantik dari sumber web. Pendekatan pertama berkaitan dengan visi web semantik yang mana data mempunyai struktur dan ontologi pula menerangkan semantik data tersebut. Namun demikian, kebanyakan dokumen web dipersembahkan dalam bentuk teks bahasa tabii dan imej yang hanya mampu dibaca dan difahami oleh manusia. Oleh yang demikian pendekatan kedua lebih praktikal untuk diaplikasikan dalam menyokong merealisasikan visi web semantik.

Penggunaan domain pengetahuan khusus dalam bentuk ontologi dilihat sebagai salah satu alternatif penyelesaian yang boleh diambil untuk mencapai dan mengekstrak kandungan maklumat semantik terutamanya bagi teks yang tidak berstruktur dalam laman web. Namun demikian, pembangunan sistem yang dapat menyokong capaian dan pengekstrakan maklumat semantik dokumen cukup mencabar disebabkan pengetahuan manusia sama ada yang tersurat mahu pun tersirat adalah berformat tekstual, tidak berstruktur dan tidak kaya dengan semantik [3].

Bahasa perwakilan web semantik adalah kompleks untuk dibangunkan, dan pengguna dijangka akan masih lagi menggunakan HTML untuk menyampaikan maklumat. Pada masa ini, terdapat lebih kurang 5 billion dokumen HTML dan bilangan ini akan semakin bertambah. Pendekatan yang sesuai untuk aktiviti capaian, pengorganisasian, pendeskripsian, pengelasan, dan persembahan maklumat dalam bentuk yang kaya semantik akan memberi manfaat yang sangat besar dalam pembangunan kandungan web semantik. Namum begitu, kebergantungan kepada pustakawan dan pakar domain akan mengambil masa yang lama dan kos yang tinggi, manakala ketiadaan pakar domain akan menjadikan hasil permodelan dokumen tidak konsisten.

Oleh sebab itu, pendekatan yang kedua dipilih kerana lebih praktikal. Sistem yang dibangunkan dapat memodelkan dan seterusnya menterjemahkan kandungan dokumen HTML kepada perwakilan yang kaya dengan semantik manakala ketiadaan pakar domain pula digantikan dengan penggunaan domain ontologi.

Pelbagai pendekatan telah dibincangkan untuk melaksanakan proses permodelan semantik kandungan dokumen web. Dua pendekatan yang menjadi pilihan adalah menggunakan pendekatan analisi bahasa tabii dan ontologi. Pendekatan ini dipilih

kerana analisis bahasa tabii diperlukan untuk menangani maklumat input kueri dan dokumen web yang kebanyakannya dikodkan dalam format teks tabii manakala ontologi pula diperlukan sebagai garis panduan untuk mendapatkan model semantik maklumat yang boleh diterima oleh semua pihak dan membantu proses capaian dan pengekstrakan maklumat. Hasilnya, gabungan antara kedua-dua pendekatan ini menyediakan antara muka capaian dan perwakilan maklumat yang lebih berkesan untuk perkongsian maklumat di web.

Kertas kerja ini menerangkan kajian penyelidikan yang dijalankan dalam membangunkan sistem capaian maklumat berasaskan semantik dan ontologi. Capaian secara semantik dokumen web lebih berkesan dengan menggunakan kaedah domain ontologi dan pemprosesan bahasa tabii kerana teknik ini juga mampu memberikan konsep yang berkaitan dengan kueri pengguna. Idea utama penyelidikan ini ialah untuk menjana model semantik dokumen setiap kali capaian dilakukan, yang mana pengguna dapat mencapai dan melayarinya secara semantik.

Kajian ini melibatkan penganalisaan ke atas kueri pengguna dan pengekstrakan maklumat semantik dokumen yang dicapai oleh enjin gelintar komersial. Hasil penganalisaan ke atas kueri pengguna adalah model semantik kueri, yang mana akan dihantar kepada enjin gelintar komersial untuk mendapatkan dokumen atas talian. Pengekstrakan maklumat semantik pula dilakukan ke atas senarai ayat terpilih bagi mendapatkan senarai konsep dan dipadankan dengan domain ontologi bagi menjana maklumat semantik kandungan dokumen. Domain ontologi yang digunakan adalah merupakan sebahagian daripada domain ontologi MeSH [4].

2.0 PERANAN ANALISIS BAHASA TABII DAN ONTOLOGI DALAM CAPAIAN MAKLUMAT SEMANTIK

Kajian telah menunjukkan bahawa terdapat pengaplikasian ontologi dan teknik pemprosesan bahasa tabii dalam capaian maklumat. Ontologi merupakan "perwakilan nyata bagi sesuatu domain"[5], yang mana konsep dan hubungannya akan diisytiharkan sebagai istilah perwakilan yang membenarkan perkongsian dan penggunaan semula maklumat [6]. Ontologi turut menyediakan konsep formal bagi sesetengah domain yang dikongsi oleh sekumpulan pengguna. Ontologi menggambarkan tatabahasa sebagai skema data kompleks yang digunakan untuk menggabungkan semantik metadata dan menawarkan nilai tambahan bagi penghuraian semantik [7]. Ontologi juga telah digunakan secara meluas dalam menyokong komunikasi dan perkongsian maklumat dalam konteks web semantik, namun potensinya dalam capaian semantik dokumen web masih belum diterokai sepenuhnya.

Ontologi memainkan peranan yang penting dalam proses perwakilan kandungan dokumen web. Memandangkan maklumat yang terkandung dalam dokumen web berada dalam format yang tidak berstruktur atau separa struktur, maka proses pengekstrakan maklumat tidak dapat dijalankan sepenuhnya berdasarkan kepada

struktur ayat. Konsep boleh dihubungkan antara satu sama lain untuk menyampaikan maklumat tanpa mengira kedudukan konsep tersebut dengan menggunakan ontologi.

Teknik pemprosesan bahasa tabii juga telah digunakan secara meluas dalam capaian maklumat disebabkan kaedah ini dapat membantu dalam meningkatkan keberkesanan proses capaian maklumat [8, 9]. Antaranya ialah bagi proses pengindeksan, pemangkatan, pengkuerian dokumen dan sebagainya. Menurut Strzalkowski [9], pernyataan ini dibuat berdasarkan andaian proses linguistik dapat merangkumi sesetengah aspek semantik yang kritikal dalam mewakilkan kandungan dokumen. Kebanyakan enjin gelintar mengaplikasikan analisis bahasa tabii dalam pemprosesan kueri pengguna [10] yang mana kaedah cantasan merupakan yang paling kerap digunakan. Proses cantasan merupakan suatu proses bagi mendapatkan kata akar sesuatu perkataan.

Pemilihan teknik pemprosesan bahasa tabii dan ontologi dalam pembangunan sistem capaian maklumat secara semantik dokumen web adalah disebabkan oleh kedua-duanya saling lengkap melengkapi yang mana teknik pemprosesan bahasa tabii diperlukan untuk menganalisis teks atau ayat dokumen web yang lengkap manakala ontologi pula diperlukan untuk menghubungkan konsep yang tidak berhubungan dalam satu ayat atau teks yang sama.

Capaian maklumat berasaskan konsep mengaplikasikan kaedah bahasa tabii dalam mengekstrak konsep yang terdapat dalam kueri pengguna dan dengan menggunakan domain ontologi, sinonim perkataan dan maksud sebenar jujukan konsep akan diberikan. Konsep yang diekstrak ini akan dipadankan dengan konsep yang diindeks. Ini dapat membantu pengguna dalam mendapatkan dokumen yang relevan dengan lebih tepat. Oleh yang demikian, penggabungan penggunaan kedua-dua teknik diharap dapat meningkatkan kualiti capaian maklumat dan menambah nilai semantik

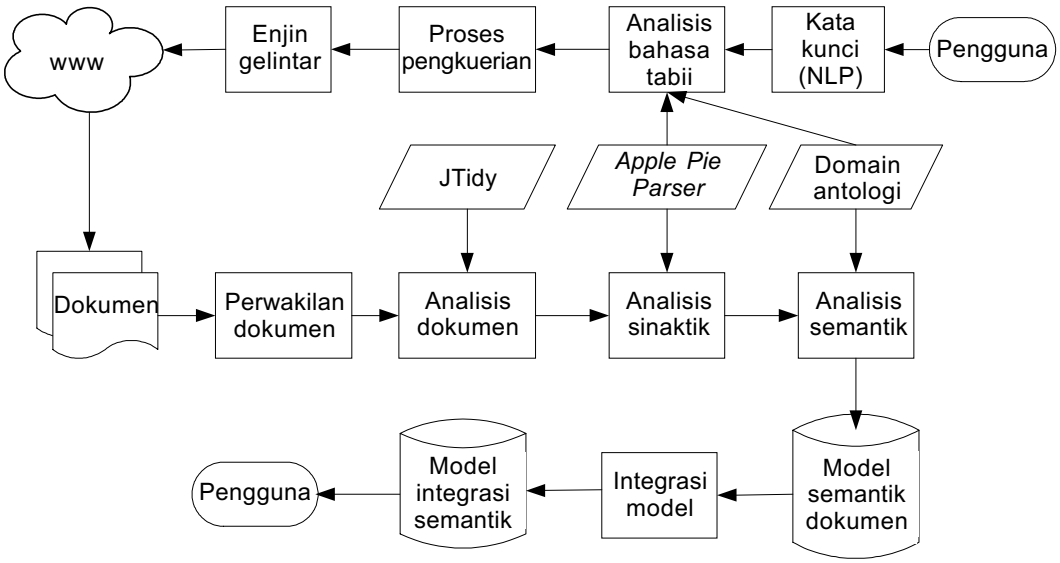
3.0 ANALISIS PENDEKATAN SEMASA

Brasethvik dan Gulla [11-13], dan Alani *et al*. [14] telah membuat kajian berkaitan penggunaan semantik dalam capaian maklumat di laman web. Brasethvik dan Gulla [11] menggunakan teknik pemprosesan bahasa tabii dan permodelan konseptual bagi proses pengklasifikasian dan capaian dokumen. Kajian ini melibatkan pengguna membangunkan model konseptualnya sendiri dengan memproses koleksi dokumen dari domain yang sama bagi mendapatkan senarai calon konsep dokumen. Model konseptual ini akan digunakan untuk permodelan, pengklasifikasian dan capaian dokumen. Bagi proses capaian dokumen, pengguna menginput kueri berbentuk bahasa tabii dan akan melalui proses linguistik bagi mendapatkan konsep model. Konsep model ini akan dipadankan dengan konsep domain model yang ada (model konseptual yang dibangunkan). Setelah konsep domain model ini ditemui, ianya akan digunakan untuk mencapai dokumen yang telah disimpan (berbentuk XML). Sistem ini menggunakan *document sevlet* sendiri untuk memproses pemadanan, penyenaraian dan persembahan dokumen.

Alani *et al*. [14] telah membangunkan sistem Artequack dengan menggunakan kaedah ontologi dalam permodelan dan capaian semantik dokumen bibliografi. Domain ontologi yang digunakan ialah berkaitan dengan artis dan artifak yang dibina berasaskan CIDOC *Conceptual Reference Model* (CRM). Teknik capaian yang digunakan ialah carian berdasarkan contoh (*searching by example*) dengan menggunakan contoh dokumen daripada laman web yang dipercayai seperti *Web Museum. Web Museum* menyediakan penerangan ringkas mengenai artis yang dicari. Gabungan penggunaan domain ontologi, WordNet [15] dan GATE (alat pemprosesan bahasa tabii bagi pengiktirafan entiti), Artequakt akan mengekstrak konsep dan hubungan di antara konsep dengan menganalisis ke semua ayat ke atas dokumen terpilih. Konsep dan hubungan serta ayat atau perenggan dalam dokumen yang menerangkan konsep dan hubungannya akan disimpan dalam bentuk fail XML (*Extensible Markup Language*).

4.0 PENDEKATAN CAPAIAN SEMANTIK DOKUMEN WEB

Kaedah yang diaplikasikan dalam pembangunan sistem secara semantik ini ialah domain ontologi dan pemprosesan bahasa tabii. Kedua-dua kaedah ini digunakan untuk analisis teks bagi input kueri dan dokumen untuk mengenal pasti konsep penting kueri dan konsep yang dapat mewakili kandungan dokumen. Ontologi juga turut digunakan untuk memberikan hubungan semantik antara konsep yang diekstrak dari proses capaian dan perwakilan dokumen. Pendekatan yang diaplikasikan ini adalah mengikut pendekatan umum dalam pembangunan indeks semantik seperti yang dinyatakan oleh Desmontils dan Jacquin [16]. Rajah 1 menunjukkan keseluruhan proses



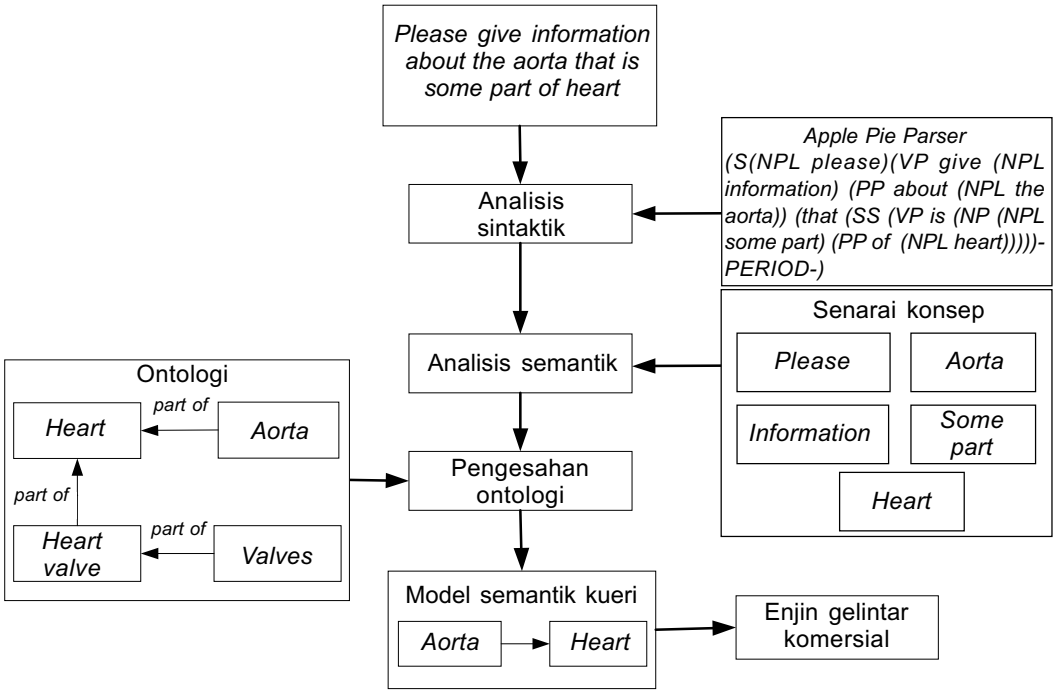
Rajah 1 Seni bina sistem capaian secara semantik dokumen web

yang terlibat dalam sistem yang dibangunkan. Jika dibandingkan dengan sistem yang dibangunkan oleh Brasethvik dan Gulla [13] yang mana domain ontologinya dibangunkan sendiri sedangkan sistem yang dibangunkan ini menggunakan domain ontologi sedia ada bagi domain pilihan iaitu domain perubatan. Setiap proses yang terlibat akan dibincangkan secara terperinci.

Secara amnya, dua proses utama yang terlibat adalah capaian dan perwakilan semantik dokumen web. Capaian merupakan suatu proses untuk mendapatkan model semantik kueri berdasarkan input kueri yang kemudiannya akan dihantar kepada enjin gelintar komersial untuk capaian dokumen web. Sementara dokumen yang dicapai akan menjalani proses perwakilan semantik dokumen yang memberikan output dalam bentuk model integrasi semantik dokumen.

4.1 Capaian

Pengguna menginput kueri berbentuk bahasa tabii dan akan melalui proses analisis bahasa tabii iaitu analisis sintaktik dan analisis semantik dalam mendapatkan model semantik kueri. Rajah 2 merupakan ilustrasi proses yang terlibat bagi mendapatkan model semantik kueri pengguna. Contoh input kueri ialah "*please give information about the aorta that is some part of heart*". Kueri yang diinput akan dihuraikan dengan menggunakan perisian linguistik sedia ada iaitu *Apple Pie Parser* (APP) [17].

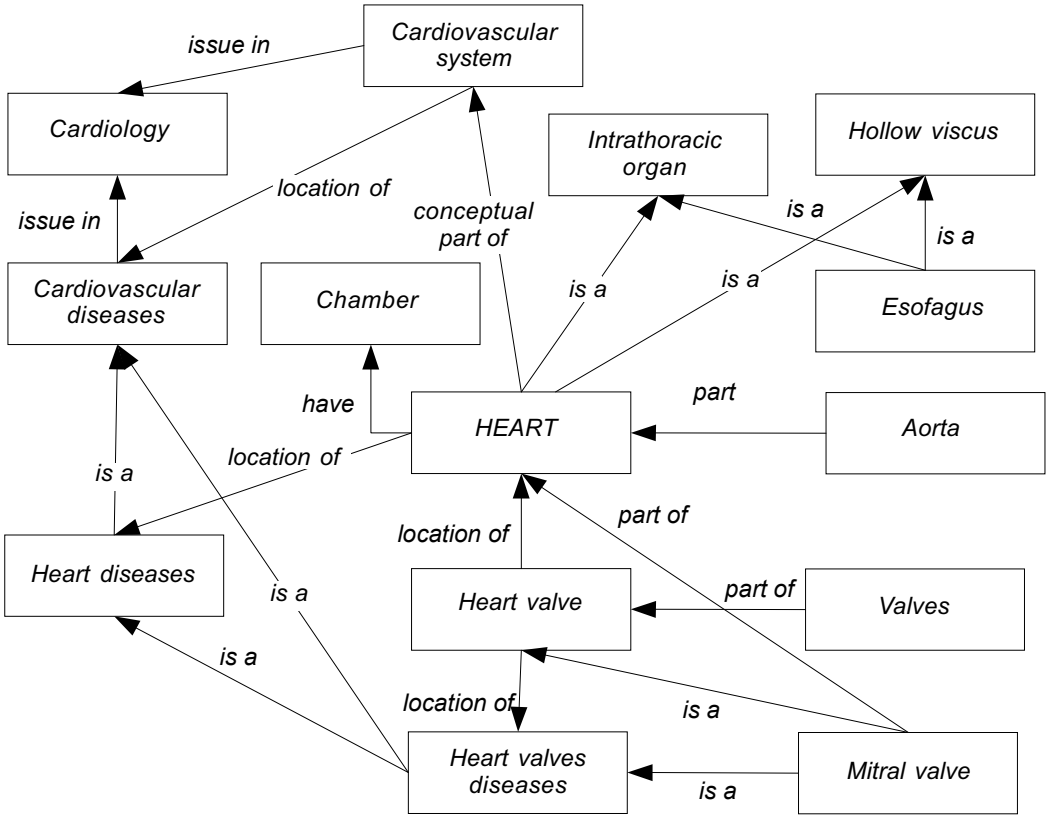


Rajah 2 Ilustrasi penjanaan model semantik kueri

Analisis semantik dilakukan dengan menggunakan output yang diberikan pada peringkat analisis sintaktik seperti Rajah 2. Setiap frasa nama akan diekstrak. Setiap frasa nama (NPL) yang diekstrak akan dianalisa bagi mencantas *determiner* (*the, a, an*) dan hasilnya adalah senarai konsep. Konsep yang disenaraikan akan dipadankan dengan domain ontologi untuk tujuan pengesahan konsep.

Domain ontologi yang digunakan untuk menguji prototaip ini ialah domain ontologi jantung seperti yang dinyatakan dalam *Medical Ontology Research* [18]. Domain ontologi ini merupakan sebahagian daripada model domain ontologi MeSH (*Medical Subject Heading*) [4]. Hubungan antara konsep akan diekstrak secara automatik berdasarkan hubungan yang terdapat dalam domain ontologi dan membentuk model semantik kueri. Rajah 3 menunjukkan sebahagian domain ontologi *heart* yang digunakan.

Model semantik kueri ini kemudiannya dihantar pada enjin gelintar komersial (MSN *search engine*) untuk mendapatkan senarai *hit* (dokumen). Keputusan capaian yang diberikan oleh enjin gelintar akan diekstrak menggunakan JTidy (*Java port of HTML Tidy*) [19] bagi mendapatkan senarai URL (*Uniform Resource Locator*) dokumen HTML



Rajah 3 Domain ontologi "*heart*"

(*Hypertext Markup Language*) dan ASP (*Active Server Page*). Dokumen yang diperoleh akan dimuat turun dan menjalani proses perwakilan dokumen bagi mendapatkan model semantik setiap dokumen.

4.2 Perwakilan Semantik Dokumen Web

Menurut Woods [20], perwakilan kandungan semantik dokumen dilakukan dengan membangunkan indeks semantik (konseptual) yang mana konsep dirangkaikan antara satu sama lain untuk memberikan penerangan maksud yang lebih jelas. Berbeza dengan indeks perkataan yang biasa, indeks semantik menghubungkan konsep atau perkataan yang terkandung dalam dokumen web. Indeks konseptual merupakan pendekatan perwakilan dokumen web dengan memodelkan kandungan maklumat secara semantik iaitu dengan menghubungkan konsep yang terdapat dalam dokumen yang dianalisis dalam bentuk model konseptual.

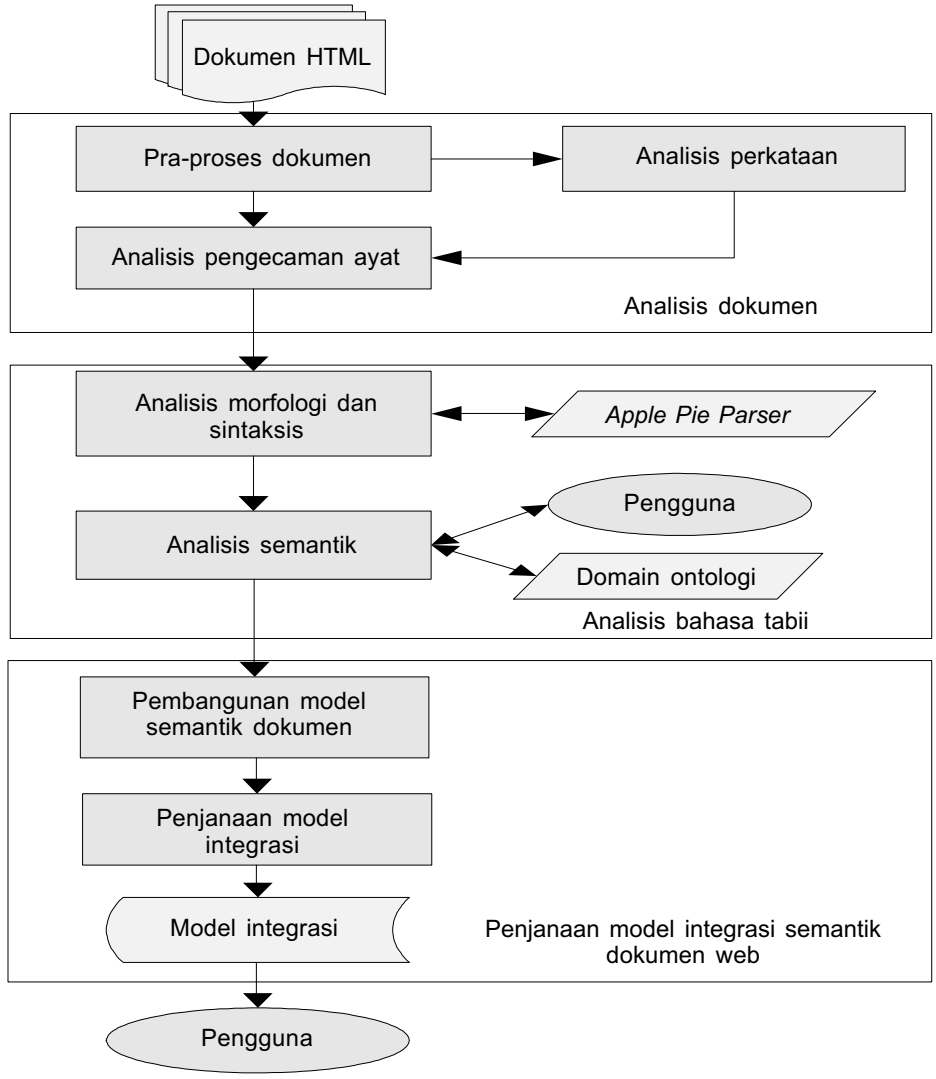
Perwakilan semantik merupakan salah satu teknik yang membenarkan sistem mengindeks kandungan maklumat menggunakan model konseptual dalam menyampaikan maklumat yang terdapat dalam dokumen web secara semantik. Indeks konseptual membolehkan sistem capaian maklumat membuat hubungan antara konsep yang diberikan oleh pengguna capaian maklumat dan menghubungkan konsep dengan konsep lain yang mungkin berkait dengan maklumat yang dikehendaki pengguna [20].

Dalam kajian penyelidikan ini, proses perwakilan dokumen ini masih menggunakan teknik pemprosesan bahasa tabii dan ontologi. Setiap dokumen diproses berasingan iaitu mengikut turutan yang disenaraikan. Setelah selesai penjanaan model semantik dokumen, model ini akan menjalani proses pengintegrasian model. Model integrasi semantik dokumen ini dapat diperincikan dan dilayari secara semantik oleh pengguna.

Rajah 4 menunjukkan turutan proses pembangunan model integrasi semantik dokumen dalam mewakilkan kandungan dokumen yang dianalisis. Dua fasa utama ialah analisis dokumen dan analisis bahasa tabii (analisis sintaktik dan semantik).

4.2.1 Analisis Dokumen

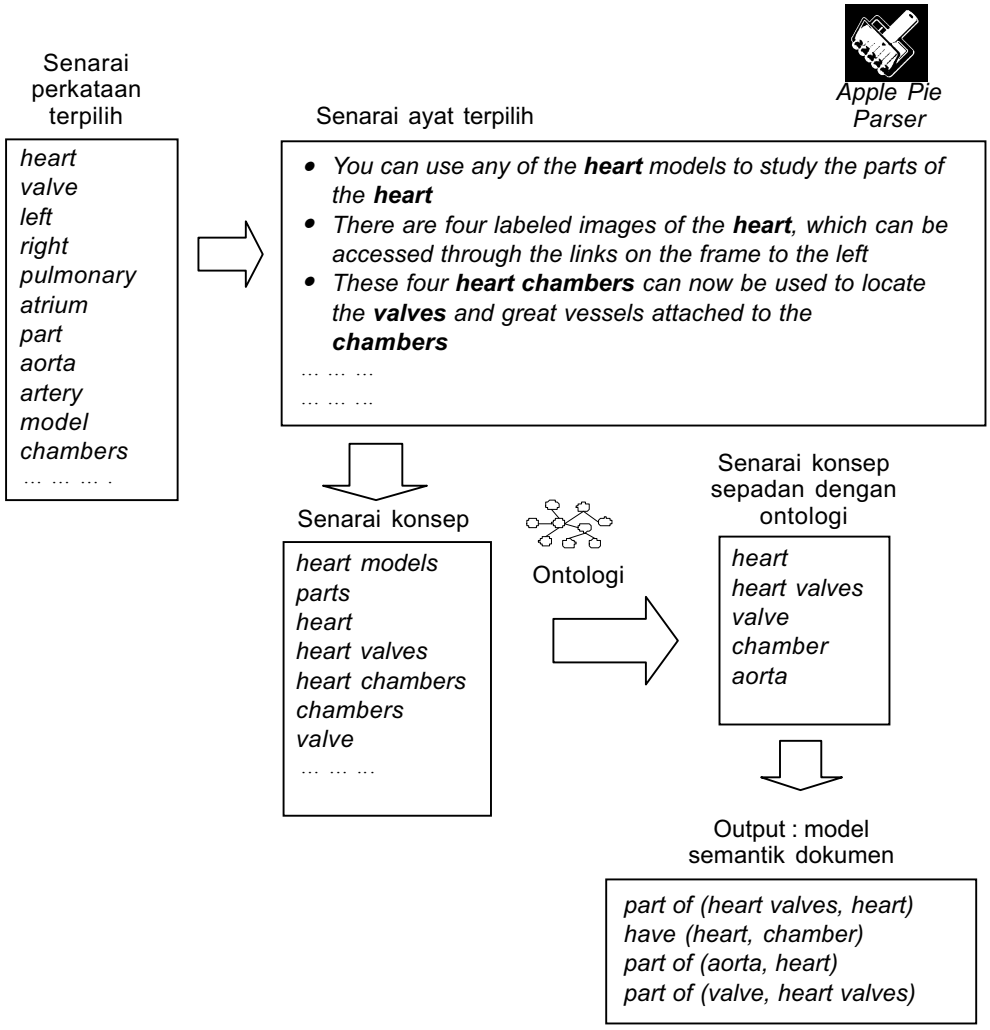
Proses perwakilan maklumat semantik dokumen web ini dilaksanakan secara automatik iaitu tidak melibatkan interaksi dengan pengguna. Dokumen HTML dan ASP akan dicapai dan dihantar kepada JTidy bagi menjalani pra-pemprosesan dokumen iaitu menukarkan dokumen tersebut ke bentuk *plain text*. Semua tag html akan dinyahkodkan menggunakan perisian Jtidy dan ianya disimpan dalam format .txt. Dokumen ini akan menjalani proses analisis kekerapan perkataan dan outputnya (senarai perkataan) akan disimpan. Senarai perkataan ini akan menjalani proses penapisan bagi menyingkirkan ke semua elemen perkataan kata henti (seperti '*the*', '*a*', '*is*', '*they*' dan sebagainya) yang tidak membawa sebarang makna dalam sebarang



Rajah 4 Pembangunan model integrasi semantik dokumen web

domain pengetahuan. Perkataan-perkataan yang tidak disingkirkan akan dipangkaskan kepada kata akar dengan menggunakan algoritma '*Poter Stemmer*'.

Sistem akan memilih perkataan yang mempunyai kekerapan tinggi sahaja untuk proses pengecaman ayat. Pemilihan perkataan ini dilakukan berdasarkan pendapat Luhn [21] yang menyatakan frekuensi data boleh digunakan untuk mengekstrak perkataan dan ayat bagi mewakili dokumen. Semasa analisis pengecaman ayat, dokumen yang dinyahkodkan akan dipecahkan kepada struktur-struktur ayat dan disimpan. Ayat yang mempunyai perkataan terpilih sahaja akan digunakan untuk proses analisis bahasa tabii.



Rajah 5 Contoh output analisis perwakilan dokumen menggunakan ontologi

4.2.2 Analisis Bahasa Tabii

Analisis ini adalah sama seperti yang diterangkan sebelum ini iaitu dalam proses mendapatkan model kueri. Ianya adalah bertujuan untuk mendapatkan model semantik bagi ayat terpilih. Keseluruhan dokumen akan menjalani proses perwakilan dokumen. Model semantik setiap dokumen akan dikumpulkan dan disimpan dalam model integrasi semantik dokumen bersama dengan alamat URL dokumen. Rajah 5 merupakan contoh output bagi setiap analisis dalam proses perwakilan dokumen. Penjanaan output ini adalah berdasarkan dokumen HTML; bio.winona.edu/dapkus/212/labs/heart1.html.

Setiap ayat yang terpilih terlebih dahulu dianalisis dalam menentukan sintaksis ayat dengan menggunakan program *Apple Pie Parser* [17]. Analisis semantik kemudian dilaksanakan bagi mendapatkan maklumat semantik yang terkandung dalam ayat terpilih. Frasa nama (*noun phrase*) yang terdapat dalam pokok huraian ayat akan diekstrak dan dianalisis bagi mencantas *determiner* (*the, a, an*). Semua konsep yang diekstrak dari ayat terpilih ini akan disenaraikan.

Proses pengekstrakan konsep ayat akan dijalankan ke atas kesemua senarai ayat terpilih. Senarai konsep keseluruhan ayat akan dipadankan dengan domain ontologi bagi mencari konsep yang sepadan dengan domain ontologi. Konsep yang sepadan dengan domain ontologi akan digunakan untuk membentuk model semantik dokumen. Hubungan antara konsep akan diekstrak secara automatik daripada domain ontologi. Output bagi proses analisis bahasa tabii ini ialah model semantik dokumen.

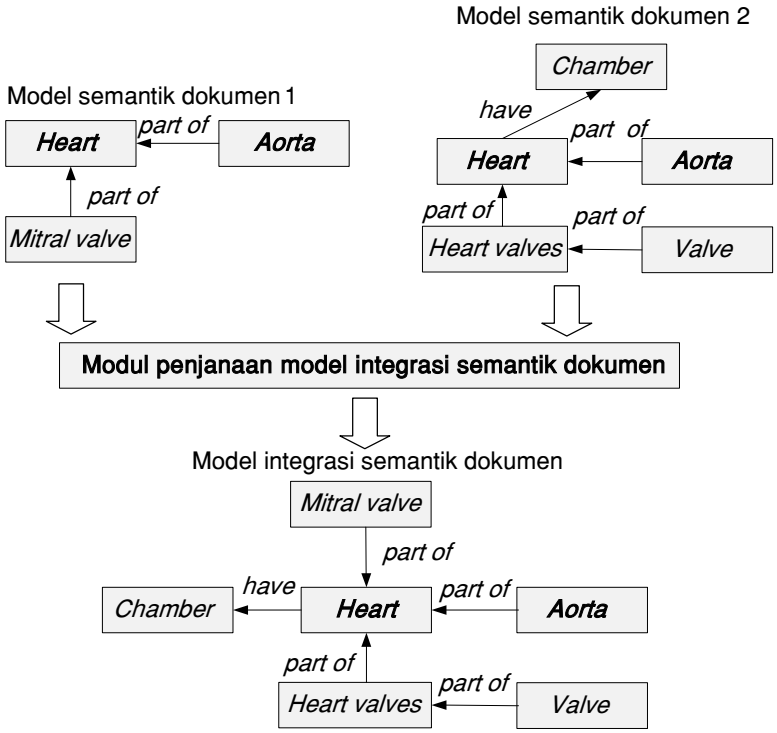
Setiap dokumen akan mempunyai model semantik kandungan dokumennya tersendiri. Kesemua model semantik dokumen ini akan dihantar untuk proses pengintegrasian model semantik dokumen. Model integrasi akan menyimpan semua model semantik dokumen bagi kesemua dokumen yang dianalisis berserta dengan senarai URL. Tujuan pembangunan model integrasi ini ialah untuk mengumpul kesemua model semantik dokumen setelah selesai proses perwakilan dokumen. Model integrasi semantik dokumen dijana dengan menggabungkan koleksi model semantik dokumen menggunakan teknik proses skema integrasi pangkalan data. Senarai berikut merupakan tiga garis panduan yang diterapkan dalam penjanaan model integrasi semantik dokumen. Dalam senarai ini, A, B dan C merujuk kepada konsep yang diuji oleh sistem.

- Jika A *part of* B dan B *part of* C, maka A *part of* C
- Jika A *is a* B dan B *is a* C, maka A *is a* C.
- Jika A *part of* B dan B *is a* C, maka A *part of* C.

Garis panduan ini penting untuk mengelakkan maklumat bakal disimpan dalam model integrasi semantik dokumen berulang (*redundent*). Rajah 6 menunjukkan contoh model integrasi bagi dua dokumen.

Merujuk kepada Rajah 6, dokumen pertama mengandungi tiga konsep iaitu "*aorta*", "*heart*" dan "*mitral valve*". Sementara dokumen kedua pula mempunyai lima konsep iaitu "*aorta*", "*heart*", "*chamber*", "*heart valves*" dan "*valve*". Oleh yang demikian konsep yang disimpan dalam model integrasi ialah "*aorta*", "*heart*", "*chamber*", "*heart valves*", "*valve*" dan "*mitral valve*". "*aorta*" dan "*heart*" akan menyimpan dua URL iaitu bagi dokumen pertama dan kedua kerana kedua-dua konsep ini terdapat dalam kedua-dua dokumen.

Model integrasi semantik dokumen dipaparkan kepada pengguna dalam bentuk hierarki. Ini bagi memudahkan pengguna untuk memahami struktur model semantik



Rajah 6 Contoh penjanaan model integrasi semantik dokumen

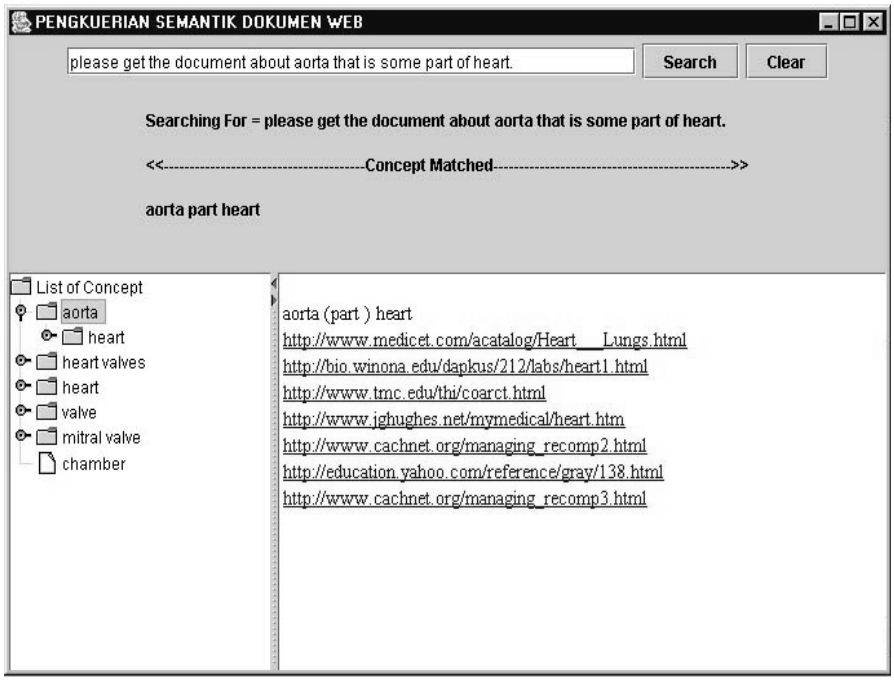
yang dijana. Model integrasi ini juga membolehkan pengguna untuk membuat perincian dan melayarinya secara semantik.

Rajah 7 menunjukkan antara muka sistem capaian yang dibangunkan. Pada antara muka ini, pengguna dapat melihat ruang untuk menginput kueri, model integrasi semantik dokumen dan senarai URL dokumen yang berkaitan dengan model integrasi pilihan. Pengguna dapat memilih model daripada model integrasi semantik dokumen bagi mendapatkan senarai URL bagi dokumen yang berkaitan. Dokumen akan dicapai dan dipaparkan kepada pengguna.

5.0 PENGUJIAN DAN PENILAIAN

Pengujian dilakukan untuk menilai keberkesanan pendekatan capaian berasaskan semantik yang dicadangkan melalui penyelidikan ini. Sehubungan dengan itu, dua bentuk pengujian dilakukan iaitu pengujian capaian semantik dokumen dan pengujian padanan konsep perwakilan semantik dokumen.

Pengujian capaian ini dilakukan untuk melihat berapakah bilangan konsep yang berkait dengan model semantik kueri pengguna dapat diberikan oleh sistem. Model semantik kueri pengguna adalah terdiri daripada dua konsep sepadan dengan domain ontologi. Pengiraan dilakukan dengan melihat jumlah konsep yang terdapat dalam



Rajah 7 Antara muka sistem capaian secara semantik

model integrasi semantik dokumen. Merujuk kepada Rajah 7, terdapat empat konsep lain iaitu *heart valves, valve, mitral valve* dan *chamber* berjaya diekstrak oleh sistem. Kesemua konsep ini mempunyai perkaitan dengan konsep kueri (*aorta, heart*).

Pengujian padanan konsep dilakukan berdasarkan teknik pengujian yang telah diaplikasikan oleh Witten *et al*. [22] dan Song *et al*. [23] dalam menguji keberkesanan sistem masing-masing iaitu KEA (*Keyphrase Extraction Analysis*) dan KPSpotter.

Kaedah yang digunakan ialah perbandingan padanan yang dibuat dengan konsep yang diekstrak oleh sistem dengan konsep dalam tag <META> kata kunci yang disediakan oleh pengarang. Witten *et al*. [22] menggunakan dokumen daripada perpustakaan digital New Zealand. Manakala Song *et al*. [23] menggunakan jurnal perubatan dalam talian. Pengujian yang dilakukan oleh Witten *et al*. [22] dan Song *et al*. [23] adalah sekadar untuk menguji konsep sepadan pada kata kuncinya sahaja. Ia berbeza dengan pengujian yang akan dilaksanakan yang mana pengujian ini dilakukan dengan menganalisis padanan konsep yang diperoleh hasil keseluruhan pemprosesan kandungan dokumen.

Sepuluh dokumen yang menyediakan tag <META> kata kunci digunakan sebagai sampel ujian perwakilan semantik dokumen. Jadual 1 menunjukkan purata padanan konsep penting dalam model semantik dokumen. Jadual ini menunjukkan bilangan konsep yang berjaya diekstrak oleh sistem iaitu sehingga enam konsep. Pengiraan purata konsep sepadan dengan tag <META> kata kunci yang disediakan oleh

Jadual 1 Purata bilangan konsep (domain perubatan)

Dokumen	**Bilangan konsep diekstrak oleh sistem**					
	1	**2**	**3**	**4**	**5**	**6**
Dok1	1	2	2	2	2	2
Dok2	0	1	1	1	1	1
Dok3	1	1	1	1	1	1
Dok4	0	1	2	2	2	2
Dok5	1	1	1	1	1	1
Dok6	0	1	1	1	1	1
Dok7	1	1	1	1	1	1
Dok8	0	1	1	1	1	1
Dok9	1	2	2	2	2	2
Dok10	1	1	1	1	1	1
Jumlah =10	6	12	13	13	13	13
	0.6	1.2	1.3	1.3	1.3	1.3

pengarang ialah nilai 1 diberikan sekiranya konsep diekstrak oleh sistem adalah sepadan dengan kata kunci pengarang dan nilai 0 sekiranya tidak sepadan.

Bagi dokumen Dok2 pula, konsep pertama bernilai kosong disebabkan konsep *aorta* tidak terdapat dalam konsep kata kunci pengarang. Konsep kedua dibandingkan dan mendapati konsep *heart* terdapat dalam kata kunci pengarang. Oleh yang demikian konsep kedua ini diberi nilai 1. Proses ini dilakukan berulang kali untuk konsep ke-n yang seterusnya.

Jika diperhatikan padanan yang dibuat adalah 'padanan tepat' sahaja iaitu tidak melibatkan kata frasa. Sehubungan dengan itu konsep '*heart*' yang terkandung dalam model semantik dokumen dan '*the heart*' atau '*heart disease*' yang terkandung dalam konsep pengarang tidak dipadankan.

Berdasarkan Jadual 2 dapat dilihat nilai purata konsep yang diekstrak oleh sistem berpadanan dengan konsep yang disediakan oleh pengarang ialah sehingga 1.3 konsep. Secara keseluruhannya, enam konsep yang berjaya diekstrak oleh sistem sekurang-

Jadual 2 Keputusan pengujian padanan konsep (domain perubatan)

Konsep yang diekstrak (sistem)	**Purata konsep sepadan**
1	0.6
2	1.2
3	1.3
4	1.3
5	1.3
6	1.3

kurangnya satu hingga dua konsep adalah sama dengan kata kunci yang disediakan oleh pengarang.

Sebagaimana yang dinyatakan pada awal perbincangan pengujian, perbandingan pengujian juga melibatkan domain pelancongan. Jadual 3 menunjukkan keputusan pengujian padanan konsep perwakilan dokumen untuk domain ini.

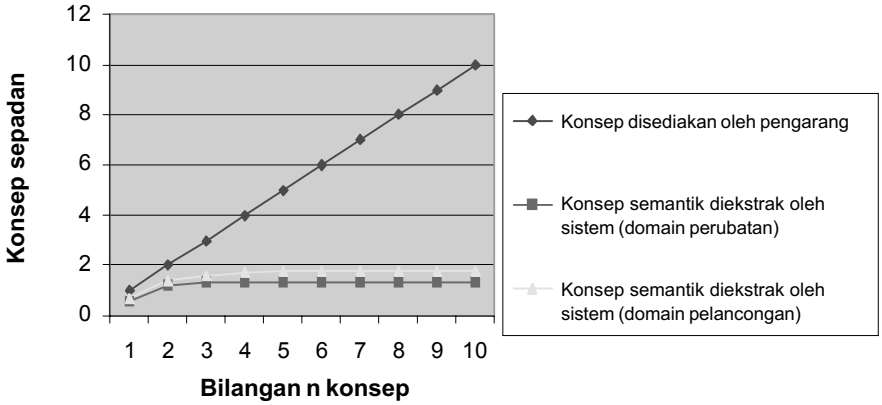
Jadual 3 Keputusan pengujian padanan konsep (domain perlancongan)

Konsep yang diekstrak (sistem)	Purata konsep sepadan
1	0.7
2	1.4
3	1.6
4	1.7
5	1.8
6	1.8
7	1.8
8	1.8
9	1.8
10	1.8

Berdasarkan Jadual 3, nilai maksimum konsep semantik yang berjaya diekstrak oleh sistem bagi satu dokumen ialah sepuluh konsep. Nilai ini lebih tinggi jika dibandingkan dengan domain perubatan (hanya enam konsep). Namun demikian, hasil analisis padanan yang diperoleh adalah hampir sama dengan hasil analisis domain perubatan. Sekurang-kurangnya satu hingga dua konsep yang diekstrak adalah sepadan dengan tag <META> kata kunci yang disediakan oleh pengarang. Daripada Jadual 2 dan Jadual 3, graf capaian semantik dokumen diplotkan dalam Rajah 8.

Hasil ini juga selari dengan hasil pengujian yang dilakukan untuk KPSpotter [23] dan KEA [22]. Keputusan ujian yang dilaksanakan untuk KPSpotter menunjukkan padanan konsep yang diperoleh adalah di antara satu hingga dua konsep (atau dengan lebih tepat 2.6 konsep). Begitu juga dengan keputusan sistem KEA yang mana hasil ujian menunjukkan padanan konsep yang diekstrak adalah antara satu hingga dua konsep (atau dengan lebih tepat 1.88 konsep).

Keputusan pengujian sistem yang dibangunkan ini boleh adalah baik memandangkan analisis perwakilan dokumen dilakukan ke atas dokumen yang tidak berstruktur di Internet. Sementara KEA melibatkan proses pengujian ke atas dokumen laporan teknikal untuk perpustakaan digital New Zealand. Manakala KPSpotter menganalisis jurnal perubatan atas talian. Kedua-dua dokumen ini adalah lebih berstruktur dan terkawal jika dibandingkan dengan dokumen web di Internet yang tiada struktur piawaian dalam perwakilan dokumen.



Rajah 8 Graf pengujian padanan konsep perwakilan semantik dokumen

Namun demikian, keputusan pengujian ini masih boleh dianggap rendah. Situasi ini berlaku disebabkan faktor berikut:

(i) Domain ontologi yang digunakan hanya merangkumi sebahagian kecil domain perubatan dan pelancongan iaitu hanya 24 konsep berkaitan dengan domain jantung digunakan sebagai domain ontologi. Di samping itu juga, konsep yang disediakan pengarang kadangkala di luar domain ontologi yang digunakan. Bilangan konsep yang dapat diekstrak dan sepadan dengan tag <META> kata kunci dijangka akan bertambah sekiranya domain ontologi yang digunakan dikembangkan lagi.

(ii) Tag <META> kata kunci yang disediakan oleh pengarang dokumen dalam pengujian pula kadangkala tidak terkandung dalam dokumen. Ini terjadi apabila dokumen pada laman web portal menggunakan satu set kata kunci untuk keseluruhan dokumen webnya. Konsep yang digunakan pula kadangkala bukan merupakan pilihan terbaik dalam menggambarkan kandungan maklumat dokumen. Contohnya turut dinyatakan nama pengarang sebagai tag <META> kata kunci dokumen web.

(iii) Kaedah padanan semasa pengujian yang hanya menggunakan padanan tepat sahaja juga menyumbang hasil padanan yang rendah.

6.0 KESIMPULAN

Sistem capaian berasaskan semantik yang dibangunkan merupakan sebagai salah satu alternatif yang boleh diambil untuk mencapai dan mengekstrak kandungan maklumat semantik bagi teks yang tidak berstruktur dalam laman web. Pendekatan

yang digunakan ialah domain ontologi dan pemprosesan bahasa tabii. Menurut Arul dan Kranti [24], pemprosesan bahasa tabii dan ontologi merupakan pendekatan terkini yang diguna pakai dalam mengenal pasti, mengekstrak dan mengorganisasi maklumat dalam dokumen web. Seterusnya maklumat ini disimpan dalam bentuk perwakilan maklumat yang lebih berstruktur dan kaya semantik

Kajian penyelidikan ini telah menunjukkan domain ontologi dan teknik analisis bahasa tabii mampu menyokong capaian semantik dokumen dengan penjanaan model semantik dokumen yang dapat diperincikan dan dilayari secara semantik. Dalam situasi ini, perincian atau pengembangan kueri pengguna boleh dicapai secara interaktif dengan cara menyusuri model integrasi semantik dokumen.

Kajian penyelidikan ini turut memperlihatkan potensi domain ontologi dalam menyokong proses perwakilan semantik kandungan dokumen web. Hasil pengujian yang dijalankan mendapati sistem yang dibangunkan ini mampu mengekstrak maklumat semantik dokumen web dan menambah nilai semantik dalam proses capaian melalui penjanaan model integrasi semantik dokumen.

Namun demikian, pengguna sistem capaian semantik ini masih memerlukan penglibatan intelektual pengguna sendiri dalam mengekspresi kueri. Oleh yang demikian, pengguna juga perlu ada literasi maklumat, dan juga pengetahuan asas dalam bahasa (dalam kajian ini bahasa Inggeris) untuk membolehkan pengguna mencapai maklumat yang dikehendaki dengan mudah.

RUJUKAN

[1] Rousseau, B. dan R. Rousseau. 2002. Some Idea Concerning the Semantic Web. Library and Information Service.

[2] van Harmelen, F. dan D. Fensel. 1999. Practical Knowledge Representation for the Web. IJCAI Workshop on Intelligent Information Integration. Stockholm, Sweden.

[3] Mattia, D. R., I. Luca, dan N. Danielle. 1998. Knowledge Representation Techniques for Information Extraction on the Web. Proceedings of the WebNet 98.

[4] Nelson, S. 2002. Medical Subject Heading (MeSH). National Library of Medicine. http://www.nlm.nih.gov/mesh/meshhome.html (dilayari pada 21 Ogos 2002)

[5] Gruber, T. A. 1999. A Translation Approach to Portable Ontology Specifications. *An International Journal of Knowledge Acquisition for Knowledge-Based Systems*. 5(2): 199-220.

[6] Villa, R., R. Wilson, dan F. Crestani. 2003. Ontology Mapping by Concept Similarity. International Conference on Digital Libraries. Kuala Lumpur, Malaysia. 666-674.

[7] Maedche, A. dan S. Staab. 2002. Applying Semantic Web Technologies for Tourism Information Systems. ENTER Conference.

[8] Croft, W. B. dan D. D. Lewis. 1987. An Approach to Natural Language for Document Retrieval. Proceedings of the 10th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 26-32.

[9] Strzalkowski, T., J. Perez-Carballo, J. Karlgren, A. Hulth, P. Tapanainen, dan T. Lahtinen. 1999. Natural Language Information Retrieval: TREC-8 Report. Proceedings of the TREC-8. 381-390.

[10] Liddy, E. 1998. Enhanced Text Retrieval Using Natural Language Processing. *Bulletin of the American Society for Information Science*. 24(4): 14-16.

[11] Brasethvik, T. dan J. A. Gulla. 1999. Semantic Accessing Document Using Conceptual Model Description. Advances in Conceptual Modeling: ER '99 Workshops on Evolution and Change in Data Management,

Reverse Engineering in Information Systems, and the World Wide Web and Conceptual Modeling. Paris, France. 321-333.

[12] Brasethvik, T. dan J. A. Gulla. 2001. Natural Language Analysis for Semantic Document Modelling. *Journal of Data and Knowledge.* 38(1): 45-62.

[13] Brasethvik, T. dan J. A. Gulla. 2002. A Conceptual Modeling Approach to Semantic Document Retrieval. Advanced Information Systems Engineering. 14th International Conference, CAiSE 2002. Ontario, Canada. 167-182.

[14] Alani, H., S. Kim, D. Millard, M. Weal, W. Hall, P. Lewis, dan N. Shadbolt. 2003. Automatic Ontology-Based Knowledge Extraction from Web Documents. *IEEE Intelligent Systems.* 18(1): 14-21.

[15] Miller, G. 1995. WordNet: A Lexical Database for English. *Communications of the ACM.* 38(11): 39-41.

[16] Desmontils, E. dan C. Jacquin. 2001. Indexing a Web Site with Terminology Oriented Ontology. International Semantic Web Working Symposiums (SWWS). Stanford University, California.

[17] Sekine, S. 2002. Proteus Project-Apple Pie Parser (*Corpus based Parser*). http://nlp.cs.nyu.edu/app (dilayari pada 15 September 2002)

[18] Bodenreider, O. 2001. Medical Ontology Research. A Report to the Board of Scientific Counselor of the Lister Hill National Centerfor Biomedical Communications. National Library of Medicine. http://lhncbc.nlm.nih.gov/cgsb/research/umls/mor (dilayari pada 20 November 2002)

[19] Sourceforge. 2003. JTidy HTML Parser and Pretty-Printer in Java. http://jtidy.sourceforge.net (dilayari pada 12 Mac 2003)

[20] Woods, W. A. 1997. Conceptual Indexing: A Better Way to Organize Knowledge. A Sun Labs Technical Report: TR-97-61 Editor. Technical Reports.

[21] Luhn, H. P. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development.* 2: 156-165.

[22] Witten, I. H., G. W. Paynter, E. Frank, C. Gutwin, dan C. G. Nevill-Manning. KEA 1999. Practical Automatic Keyphrase Extraction. Proceeding DL '99. 254-256.

[23] Song, M., I-Y. Song, dan X. Hu. 2004. An Efficient Keyphrase Extraction System Using Data Mining and Natural Language Processing Techniques. First International Workshop on Semantic Web Mining and Reasoning (SWMR 2004). Beijing, China. 58-65.

[24] Arul P. A. dan K. Kranthi. 2001. Web Page Classification based on Document Structure. IEEE National Convention.