Jurnal Teknologi Maklumat & Multimedia **4**(2007): 1-18

Pengekstrakan dan Perwakilan Semantik Dokumen Web Berorientasikan Domain Ontologi

ARIFAH CHE ALHADI, LAILATUL QADRI BINTI ZAKARIA,
SHAHRUL AZMAN MOHD NOAH, TENGKU MOHD TENGKU SEMBOK

ABSTRAK

Internet menjadi pilihan sebagai prasarana asas bagi mendapatkan maklumat digital pelbagai topik dari seluruh dunia. Namun demikian kebanyakan dokumen web dalam Internet ini adalah tidak berstruktur dan tidak mempunyai maklumat semantik dokumen. Sistem pengekstrakan maklumat yang ada lebih memfokuskan kepada pengekstrakan konsep penting dalam mewakili kandungan dokumen tanpa mengambil kira aspek semantik. Perwakilan kandungan maklumat dalam bentuk kaya semantik merupakan salah satu visi web semantik. Kertas ini membincangkan pengaplikasian pendekatan ontologi dan pemprosesan bahasa tabii dalam menyokong pengekstrakan dan perwakilan maklumat semantik dokumen web. Memandangkan penganotasian maklumat semantik secara manual daripada dokumen web adalah tidak praktikal dan pembangunan sistem automatik sepenuhnya masih terlalu awal untuk diimplementasikan, maka pendekatan separa-automatik telah diusulkan. Dalam hal ini, sistem berfungsi untuk memandu pengguna dalam pemodelan semantik dokumen web yang seterusnya menghasilkan kandungan dokumen web atau set dokumen web yang lebih kaya semantik. Model semantik yang dijana diwakilkan dalam format XML.

Katakunci: Perwakilan semantik dokumen, pengekstrakan maklumat semantik, ontologi, analisis bahasa tabii.

ABSTRACT

Internet has been chosen as a basic infrastructure to gain various topics of digital information from all over the world. However, most of the web documents are unstructured and lack of semantics. Existing information extraction system mainly concerns with extracting important keywords or key phrases that represent the content of the documents without considering the semantic aspects. The semantic representation of documents currently forms the vision of semantic web. In this paper, we discuss an approach meant to assist in extracting and modeling the semantic information content of web

documents using natural language analysis technique and a domain specific ontology. As the manual semantic annotation of web documents is impractical and unscalable; and fully automated tools are still at the very early stage to be implemented, we proposed a semi-automatic approach. In this situation, the system will guide the user in semantic document modeling which results in the generation of semantic-rich web document content represented as XML.

Keywords: Semantic document representation, semantic information extraction, ontology, natural language analysis.

PENGENALAN DAN PERNYATAAN MASALAH

Hypertext Markup Language (HTML) telah digunakan secara meluas sebagai kod piawai dalam penyebaran maklumat dalam web. Maklumat boleh dikodkan dengan mudah, menggunakan kod HTML. Gabungan antara HTML dan *Hypertext Transport Protocol* (HTTP) telah membawa kepada perubahan besar dalam cara manusia menghantar dan menerima maklumat digital. Bagaimanapun, dokumen HTML hanya menyediakan kemudahan untuk memaparkan maklumat dan bukan untuk mewakilkan atau memodelkan maklumat. Ini menyebabkan HTML hanya memfokuskan kepada bagaimana sesuatu maklumat dipersembahkan kepada pengguna dan bukannya pada maklumat itu sendiri atau bagaimana untuk menstrukturkan maklumat (Harmelen & Fensel 1999).

Sehubungan dengan itu, pengetahuan manusia yang tersurat dan tersirat dalam Web masih lagi dalam keadaan tidak berstruktur dan tidak kaya semantik (Rosa et al. 1998). Aspek ini mengundang ketidakmampuan bagi mesin atau program untuk mendeduksi pengetahuan yang terkandung dalam korpus maklumat yang besar ini.

Sebagaimana yang dinyatakan oleh Zadeh (2004), pengkuerian pengetahuan seperti "*Dapatkan masa operasi Hospital UKM pada hari minggu*" tidak akan sama sekali dapat dilakukan dalam korpus maklumat web masa kini. Usaha awal yang dilakukan ialah dengan menstrukturkan kandungan dokumen web supaya lebih kaya semantik sebagaimana yang diusulkan dalam visi Semantic Web (Berners-Lee et al. 2001), Cyc (Lenat 1995), OWL (Smith et al. 2003) dan sistem lain yang berdasarkan ontologi (Smith & Welty 2002; Smith et al. 2003 & Sowa 1999). Walau bagaimanapun anotasi semantik yang diperkenalkan oleh Semantic Web, Cyc dan OWL masih lagi kompleks untuk dipraktikkan dan dengan lebih 1.5 bilion dokumen dalam bentuk HTML, maka bentuk perwakilan HTML masih lagi menjadi pilihan pembangun dan pengarang.

Penganotasian maklumat semantik secara manual daripada dokumen web adalah tidak praktikal dan pembangunan sistem automatik sepenuhnya masih terlalu awal untuk diimplementasikan (Rousseau & Rousseau 2002). Sehubungan dengan itu pendekatan separa-automatik adalah lebih praktikal

yang berfungsi untuk memandu pengguna dalam pemodelan semantik dokumen web yang seterusnya menghasilkan kandungan dokumen web atau set dokumen web yang lebih kaya semantik.

Kertas ini membincangkan penggunaan teknik pemprosesan bahasa tabii dan domain ontologi dalam membangunkan sebuah sistem yang berperanan untuk menguruskan maklumat semantik yang tersurat dalam dokumen web HTML secara sistematik serta menyimpan maklumat semantik ini dalam format yang menyokong persembahan dan perwakilan maklumat dalam WWW. Analisis atau teknik pemprosesan bahasa tabii telah banyak digunakan dalam sistem pemprosesan teks seperti sistem pengekstrakan maklumat, aplikasi untuk meringkaskan teks dan sistem perlombongan teks. Sebagaimana yang akan dibincangkan dalam bahagian seterusnya, teknik pemprosesan bahasa tabii dilihat berpotensi untuk membantu dalam proses perwakilan semantik dokumen web. Umumnya ontologi ditakrifkan sebagai spesifikasi untuk suatu pengkonseptualan (Gruber 1993). Ontologi bukanlah sesuatu yang baru dalam konteks perwakilan pengetahuan, bagaimanapun penggunaannya hanya menjadi semakin popular dewasa ini dengan perkembangan web semantik. Dalam konteks web semantik, ontologi banyak digunakan dalam aspek pengintegrasian dan pemetaan maklumat dan pengetahuan. Seperti mana teknik pemprosesan bahasa tabii, ontologi juga dilihat dapat membantu dalam pengekstrakkan maklumat semantik dokumen web.

ANALISIS PENDEKATAN SEMASA

Pengekstrakan kandungan laman web adalah bertujuan untuk mendapatkan senarai kata kunci yang relevan dan mencerminkan kandungan sesuatu laman web. Pengekstrakan maklumat ini melibatkan proses mengenalpasti bahagian teks yang relevan, pengekstrakan maklumat yang releven dari bahagian teks yang dikenalpasti dan menyimpan maklumat tersebut dalam bentuk struktur rangkaian yang mempunyai nilai semantik. Pelbagai teknik telah diperkenalkan bagi mengenalpasti maklumat yang terdapat dalam dokumen web HTML, mengekstrak dan mengorganisasi maklumat tersebut dan seterusnya menyimpan maklumat dalam perwakilan yang lebih berstruktur dan kaya semantik (Arul & Kranthi 2001). Masalah pengekstrakan maklumat ini seringkali dikaitkan dengan teknik pemprosesan bahasa tabii dan ontologi. Ontologi merupakan "perwakilan nyata bagi sesuatu domain" (Gruber 1999), yang mana konsep dan hubungannya akan diisytiharkan sebagai istilah perwakilan yang membenarkan perkongsian dan penggunaan semula maklumat (Villa et al. 2003).

Beberapa kajian dalam pengekstrakan maklumat dan permodelan semantik dokumen web boleh dilihat pada hasil kajian yang dilakukan oleh Brasethvik dan Gulla (2001) dan Alani et al. (2003). Kedua-dua kajian ini melibatkan pengaplikasian pendekatan pemprosesan bahasa tabii dan ontologi. Brasethvik

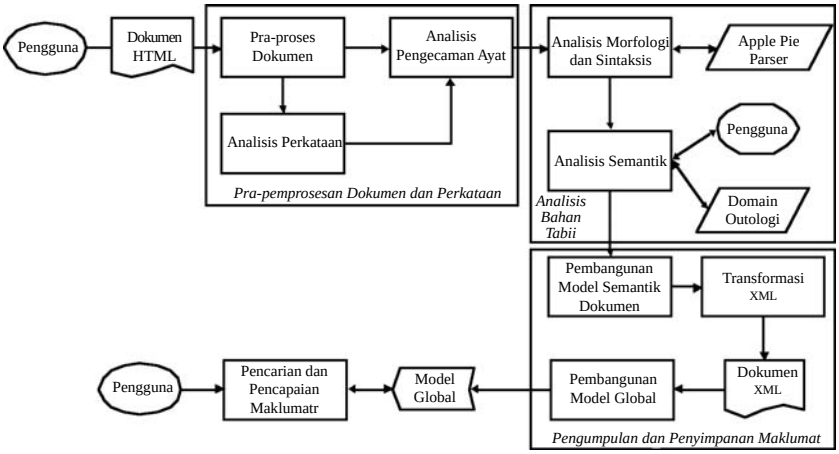
dan Gulla (2001) menggunakan kaedah pemprosesan bahasa tabii dan permodelan konseptual bagi proses pengklasifikasian dan capaian dokumen. Kajian ini melibatkan pengguna membangunkan model konseptualnya sendiri dengan memproses koleksi dokumen dari domain yang sama bagi mendapatkan senarai calon konsep 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. (2003) telah membangunkan sistem Artequack dengan menggunakan kaedah ontologi dalam permodelan dan capaian semantik dokumen bibliografi. Domain ontologi yang digunakan ialah berkaitan dengan pelukis 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 [11] 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*).

PENDEKATAN PENGEKSTRAKAN DAN PENGINTEGRASIAN MAKLUMAT SEMANTIK DOKUMEN

Kaedah yang diaplikasikan dalam pembangunan sistem pengekstrakan dan permodelan semantik dokumen ini ialah domain ontologi khusus dan pemprosesan bahasa tabii. Kedua-dua kaedah ini digunakan untuk analisis teks bagi dokumen untuk mengenal pasti konsep penting dan hubungan antara konsep tersebut yang dapat mewakili kandungan semantik dokumen. Penggunaan domain pengetahuan khusus dalam bentuk ontologi dilihat sebagai salah satu alternatif penyelesaian yang boleh diambil bersesuaian untuk jangka masa singkat ini untuk mengekstrak kandungan maklumat semantik terutamanya bagi teks yang tidak berstruktur dalam laman web. Pendekatan yang diaplikasikan ini adalah mengikut pendekatan umum dalam pembangunan indeks semantik seperti yang dinyatakan oleh Desmontils dan Jacquin (2001). Sistem ini dibangunkan untuk membolehkan pengguna (pakar domain)

mengenalpasti maklumat penting yang mewakili dokumen dan seterusnya disimpan dalam bentuk perwakilan yang lebih kaya semantik. Rajah 1 menunjukkan keseluruhan proses yang terlibat dalam sistem yang dibangunkan. Jika dibandingkan dengan sistem yang dibangunkan oleh Brasethvik dan Gulla (2001) yang mana domain ontologinya dibangunkan sendiri sedangkan sistem yang dibangunkan ini menggunakan domain ontologi sedia ada bagi domain pilihan iaitu domain perubatan. Sistem ini terbahagi kepada empat fasa utama iaitu fasa prapemprosesan dokumen dan perkataan, analisis bahasa tabii, penyimpanan dan pengumpulan maklumat dan carian maklumat. Setiap proses yang terlibat akan dibincangkan secara terperinci.

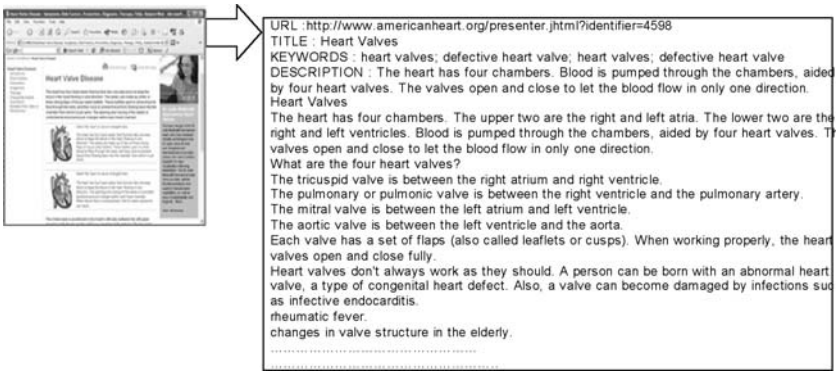


RAJAH 1. Proses sistem pemodelan semantik kandungan dokumen web

PRAPEMPROSESAN DOKUMEN DAN PERKATAAN

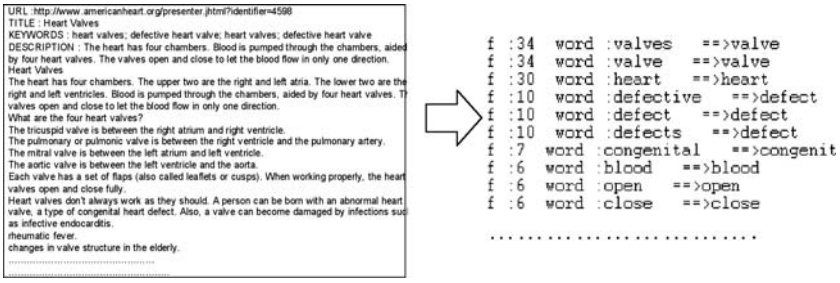
Dokumen HTML akan menjalani pra-pemprosesan dokumen iaitu menukarkan dokumen tersebut ke bentuk *plain text.* Semua tag html akan dinyahkodkan menggunakan perisian *Jtidy* dan ianya disimpan dalam format .txt. Rajah 2 menujukkan contoh output dokumen yang dijana daripada proses pemprosesan dokumen. Dokumen 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 domain pengetahuan. Perkataan-perkataan yang tidak disingkirkan akan dipangkaskan kepada kata akar dengan menggunakan algoritma '*Porter Stemmer'*. Contohnya, perkataan 'valves' akan dipangkaskan menjadi perkataan *'valve'*.

Sistem akan memilih perkataan yang mempunyai kekerapan tinggi sahaja berdasarkan kekerapan kata dasar untuk proses pengecaman ayat. Pemilihan



RAJAH 2. Prapemproses dokumen HTML kepada teks *plain*

perkataan ini dilakukan berdasarkan pendapat Luhn (1998) yang menyatakan frekuensi data boleh digunakan untuk mengekstrak perkataan dan ayat bagi mewakili dokumen. Frekuensi perkataan yang wujud dalam dokumen yang dianalisis merupakan ukuran signifikan perkataan yang penting manakala senarai frekuensi tertinggi merupakan *hint* utama kandungan dokumen. Rajah 3 menunjukkan contoh output bagi analisis frekuensi perkataan.



RAJAH 3. Analisis frekuensi perkataan

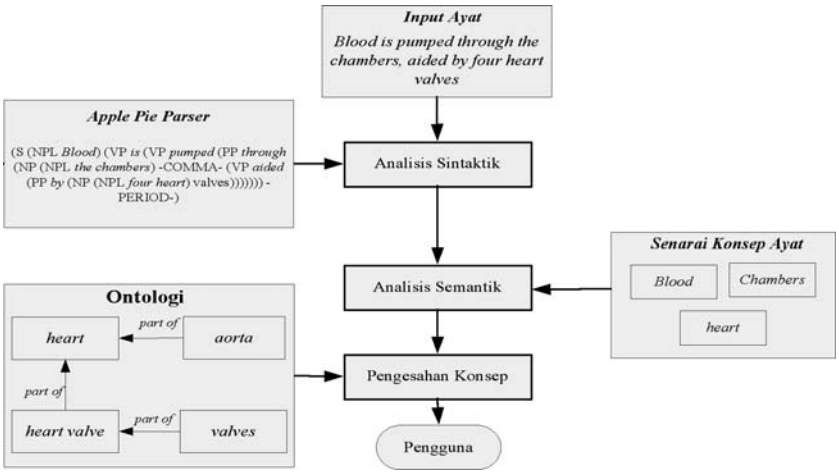
Semasa analisis pengecaman ayat, dokumen yang dinyahkodkan akan dipecahkan kepada struktur-struktur ayat dan disimpan. Ayat yang mempunyai perkataan terpilih sahaja akan digunakan untuk proses analisis bahasa tabii.

ANALISIS BAHASA TABII

Analisis bahasa tabii dijalankan bagi mendapatkan model semantik bagi mewakili kandungan maklumat dokumen web. Model semantik setiap dokumen yang dianalisis akan dikumpulkan dan disimpan dalam model global semantik dokumen bersama dengan alamat URL dokumen. Analisis bahasa tabii

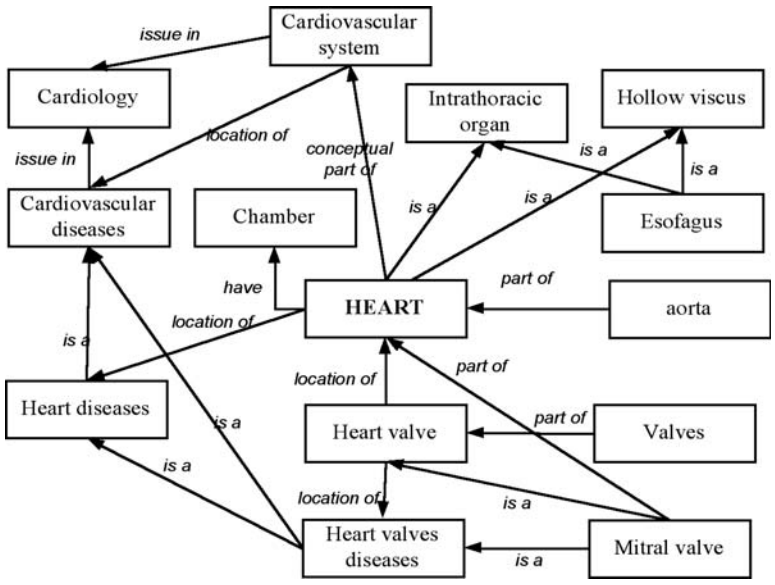
dibahagikan kepada dua peringkat iaitu analisis morfologi dan analisis sintaksis dan analisis semantik. Analisis morfologi dan sintaksis dilakukan dengan berbantukan perisian linguistik sedia ada iaitu Apple Pie Parser (APP) (Sekine 2002). Rajah 4 menunjukkan illustrasi bagi analisis sintaktik dan semantik bagi input ayat "*Blood is pumped through the chambers, aided by four heart valves"*. Setiap ayat atau frasa yang diinputkan ke dalam APP akan dihurai menjadi pohon huraian (*parse tree*). Analisis yang dilaksanakan oleh APP adalah bersifat bebas domain. Ini membolehkan penghuraian dapat dilaksanakan pada mana-mana ayat atau bahagian teks tanpa perlu merujuk kepada domain ontologi yang diwakili oleh teks tersebut.

Analisis semantik pula dilakukan dengan menggunakan output yang diberikan pada peringkat analisis sintaktik. Setiap frasa nama (NPL) yang diekstrak akan dianalisis bagi mencantas *determiner (the, a, an)* dan hasilnya adalah senarai konsep. Konsep yang disenaraikan akan dipadankan dengan domain ontologi untuk tujuan pengesahan konsep. Analisis semantik ini dijalankan bertujuan untuk mengekstrak maklumat semantik yang terkandung dalam setiap ayat terpilih dengan mengekstrak konsep dan hubungan di antara konsep yang telah dikenalpasti. Hubungan antara konsep ini boleh dilakukan dengan mengekstrak secara automatik hubungan antara konsep yang terdapat dalam domain ontologi ataupun dengan meneliti hubungan yang terdapat dalam struktur ayat. Penilaian hubungan berdasarkan struktur ayat ini memerlukan penglibatan pengguna dalam menentukan hubungan antara konsep yang berjaya diesktrak.



RAJAH 4. Ilustrasi analisis sintaktik dan semantik

Domain ontologi yang digunakan adalah merupakan domain ontologi jantung seperti yang dinyatakan dalam *Medical Ontology Research* (Bodenreider 2001). Domain ontologi ini merupakan sebahagian daripada model domain ontologi MeSH (*Medical Subject Heading*). Rajah 5 menunjukkan sebahagian domain ontologi *heart* yang digunakan.

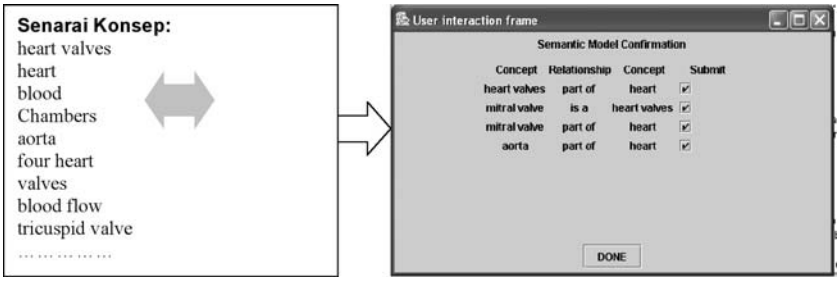


RAJAH 5. Domain ontologi "*heart*"

Analisis semantik dilaksanakan secara secara automatik dan dengan bantuan pengguna sistem. Namun demikian, pada peringkat pemodelan semantik secara automatik, pengguna masih lagi memainkan peranan dalam menentukan kesahihan model maklumat yang diekstrak oleh sistem. Analisis semantik ini tidak dapat dilaksanakan secara automatik sepenuhnya memandangkan keperluan kepada pemahaman dan pengetahuan yang luas tentang struktur dan kandungan maklumat yang hendak disampaikan (Snoussi et al. 2002).

Pemodelan Maklumat Secara Automatik. Proses permodelan maklumat semantik dijalankan dengan menganalisis 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 (Rajah 6). Hubungan antara konsep akan diekstrak secara automatik daripada domain ontologi. Pada peringkat ini pengguna terlibat

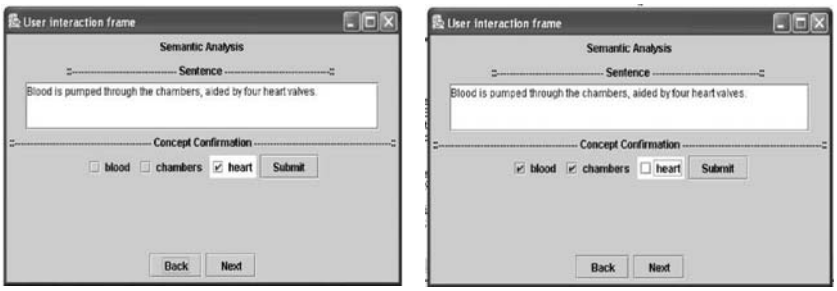
untuk membuat pengesahan bagi memastikan maklumat semantik yang diesktrak oleh sistem mencerminkan kandungan sebenar maklumat dokumen web yang dianalisis.



RAJAH 6. Pengekstrakan model semantik maklumat secara automatik

Pemodelan Maklumat dengan Berbantukan Pengguna. Sistem hanya akan memaparkan kesemua konsep dan hubungan antara konsep yang sepadan dengan domain ontologi. Namun demikian tidak semua konsep yang diberikan sesuai untuk digunakan dalam pemodelan semantik kandungan dokumen web. Pengguna berkuasa mutlak dalam menentukan konsep yang difikirkan relevan menggambarkan kandungan maklumat semantik dokumen. Setiap ayat yang dianalisis akan dipaparkan kepada pengguna melalui antara muka interaksi pengguna, khusus untuk analisis semantik.

Rajah 7 menunjukkan turutan pengesahan konsep yang perlu dilakukan oleh pengguna sistem untuk mendapatkan model maklumat yang terdapat dalam ayat terpilih. Konsep *heart* merupakan konsep sepadan dengan domain ontologi yang digunakan. Sementara *blood* dan *chambers* merupakan pilihan yang dibuat oleh pengguna yang merasakan kedua-dua konsep ini penting dalam menggambarkan kandungan dokumen web yang dianalisis.

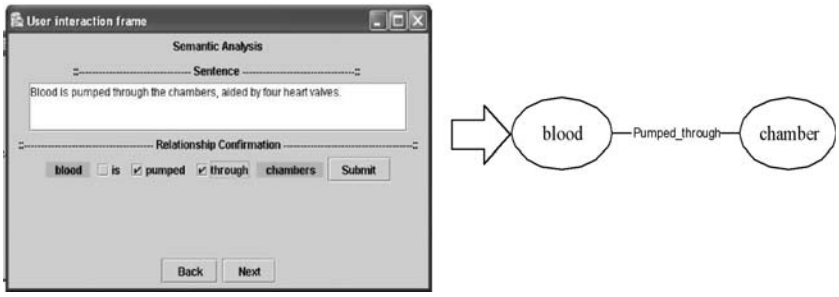


1. Cadangan konsep oleh sistem

2. Pengesahan konsep yang dilakukan oleh pengguna

RAJAH 7. Pengecaman dan pengesahan konsep berbantukan pengguna

Proses pengecaman hubungan antara konsep lazimnya dilakukan dengan menganalisis kata kerja (*verb phrase*) yang terdapat dalam struktur ayat. Sistem akan mengekstrak kata kerja dan perkataan lain yang terdapat di antara konsep pilihan pengguna (*blood* dan *chambers*) dan mencadangkannya sebagai hubungan semantik. Pemilihan hubungan di antara dua konsep ini bergantung kepada pemahaman pengguna dan kesesuaian perkataan tersebut menjadi penghubung kepada konsep '*blood*' dan '*chambers*'. Berdasarkan kesesuaian perkataan dan pemahaman, pengguna mungkin memilih perkataan '*pumped*' dan '*through*' sebagai hubungan kepada konsep '*blood*' dan '*chambers*' (Rajah 8). Oleh itu, model maklumat yang diperolehi hasil daripada analisis ayat '*Blood is pumped through the chambers, aided by four heart valves*' adalah *pumped_through (blood, chambers)*.

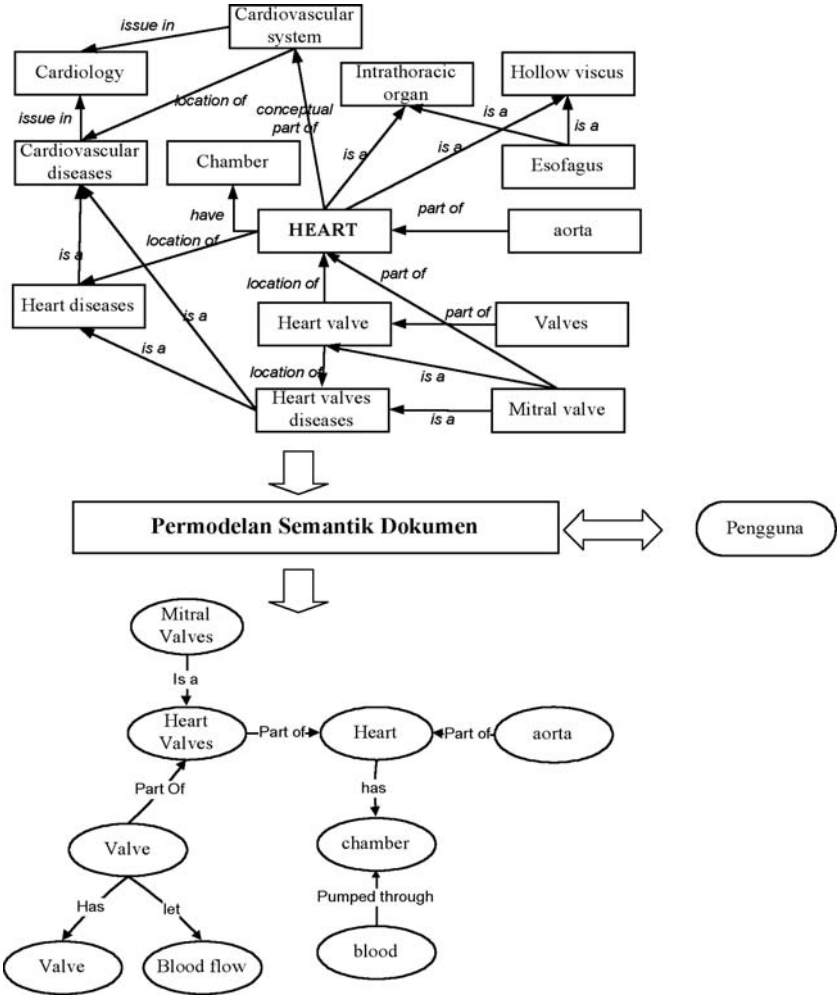


RAJAH 8. Pengesahan hubungan antara konsep '*blood*' dan '*chambers*'

Rajah 9 pula menunjukkan model semantik dokumen web yang dijana bagi satu dokumen web pilihan. Seperti yang dapat dilihat berdasarkan rajah ini, hubungan semantik bagi *"mitral valve part-of heart", "heart valve part-of heart"* dan *"mitral valve is-a heart valve"* diekstrak secara automatik daripada domain ontologi dan selebihnya merupakan konsep dan hubungan semantik yang telah ditentukan oleh pengguna berdasarkan analisis ke atas struktur ayat. Model semantik dokumen web ini akan disimpan dalam bentuk XML berserta dengan URL dokumen tersebut.

PENGINTEGRASIAN MAKLUMAT SEMANTIK DOKUMEN

Model global akan menyimpan semua model semantik dokumen bagi kesemua dokumen yang dianalisis berserta dengan senarai URL. Tujuan pembangunan model global ini ialah untuk mengumpul kesemua model semantik dokumen setelah selesai proses perwakilan dokumen. Model global 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 global semantik



RAJAH 9. Ilustrasi penjanaan model semantik dokumen

dokumen. Dalam senarai ini, A, B dan C merujuk kepada konsep yang diuji oleh sistem.

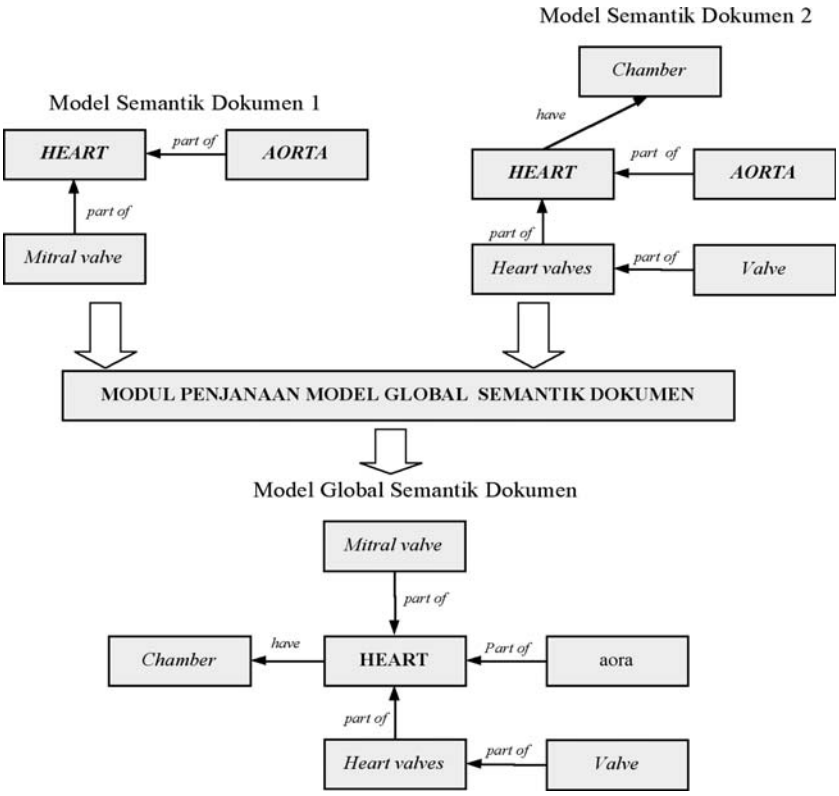
- Jika A *part of* B dan B *part of* C, maka A *part of* C
- Jika A *is a* B dan B *is a* C, maka A *is a* C.
- Jika A *part of* B dan B *is a* C, maka A *part of* C.

Garis panduan ini penting untuk mengelakkan maklumat bakal disimpan dalam model global semantik dokumen berulang (*redundent*). Rajah 10 menunjukkan contoh model global bagi dua dokumen.

Berikut merupakan struktur dokumen XML yang dijana berdasarkan maklumat yang terkumpul dalam 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.

```
<?xml version=”1.0" encoding=”UTF-8"?>
<DocumentInfo>
    <MetadataInfo>
        <Title>Heart Valves</Title>
        <Url>http://www.americanheart.org/presenter.jhtml?</Url>
        <Keywords> heart valves , heart , valve , chamber , aorta , blood
,           chambers , valves , blood flow , flaps , </Keywords>
</MetadataInfo>
<Semantic_Content>
        <Concept>
        <ConceptDescription>
        <String> heart valves </String>
        </ConceptDescription>
        <ConceptRelationship>
            <partof>
            <String>heart </String>
            </partof>
            </ConceptRelationship>
        </Concept>
        <Concept>
        <ConceptDescription>
        <String> valve </String>
        </ConceptDescription>
        <ConceptRelationship>
            <partof>
            <String> heart valves </String>
            </partof>
            <has>
            <String> flaps </String>
            </has>
        </ConceptRelationship>
        </Concept>
        </ Semantic_Content >
```

Merujuk kepada Rajah 10, dokumen pertama mengandungi tiga konsep iaitu *“aorta”*, *“heart”* dan *“mitral valve”*. Sementara dokumen kedua pula mempunyai lima konsep iaitu *“aorta”*, *“heart”*, *“chamber”*, *“heart valves”*

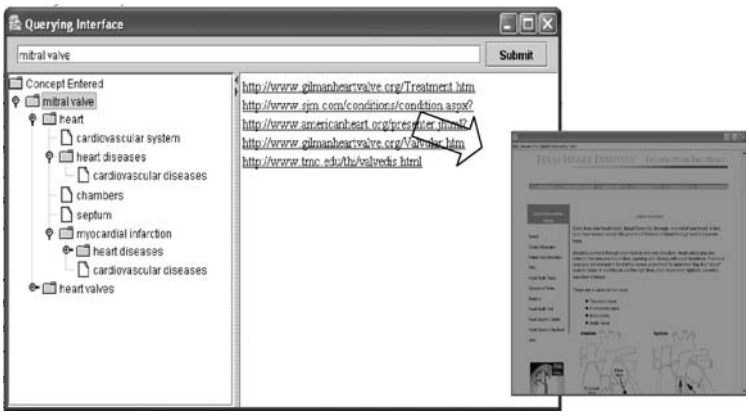


RAJAH 10. Contoh penjanaan model global semantik dokumen

dan *"valve"*. Oleh yang demikian konsep yang disimpan dalam model global 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 global semantik dokumen dipaparkan kepada pengguna dalam bentuk hierarki. Ini bagi memudahkan pengguna untuk memahami struktur model semantik yang dijana.

CARIAN DAN CAPAIAN DOKUMEN

Fasa pencarian dan capaian maklumat menyediakan antara muka khusus untuk pengguna membuat pengkuerian bagi mendapatkan semula maklumat yang telah disimpan. Rajah 11 menunjukkan hasil carian untuk konsep '*mitral valve*'. Konsep yang mempunyai hubungan dengan *mitral valve* akan dipaparkan dalam bentuk hierarki bagi membolehkan pengguna untuk membuat perincian dan melayarinya secara semantik. Pada antara muka ini, pengguna



RAJAH 11. Paparan carian dan capaian maklumat '*mitral valve*'

disediakan ruang untuk menginput kueri, ruang paparan model global semantik dokumen dan senarai URL dokumen yang berkaitan dengan konsep kueri.

PENGUJIAN SISTEM

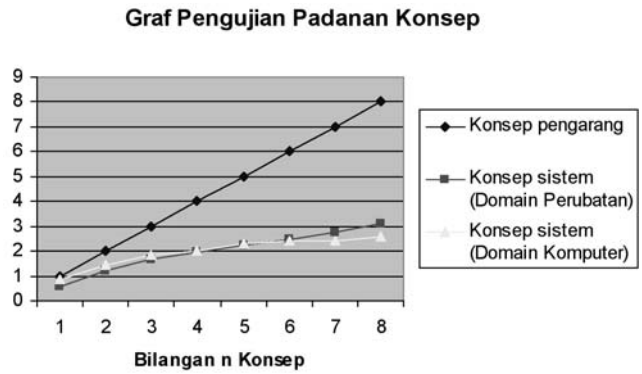
Pengujian dilakukan untuk menilai keberkesanan pendekatan permodelan semantic dokumen yang dicadangkan melalui penyelidikan ini. Pengujian padanan konsep dilakukan berdasarkan teknik pengujian yang telah diaplikasikan oleh Witten et al. (1999) dan Song et al. (2004) dalam menguji keberkesanan sistem masing-masing iaitu KEA (*Keyphrase Extraction Analysis*) dan KPSpotter. Untuk menguji keberkesanan sistem yang dibangunkan, sebanyak 50 dokumen HTML dalam talian dari domain perubatan telah dipilih secara rawak. Pemilihan dokumen HTML dilaksanakan berdasarkan kriteria berikut:

- dokumen mengandungi tag <META> kata kunci yang disediakan oleh pengarang
- domain pengetahuan dokumen adalah perubatan yang mengkhususkan kepada konsep jantung dan konsep-konsep lain yang berkaitan dengannya.

Kaedah yang digunakan ialah dengan membuat perbandingan padanan dengan konsep yang diekstrak oleh sistem dengan konsep dalam tag <META> kata kunci yang disediakan oleh pengarang. Jadual 1 menunjukkan keputusan yang diperoleh hasil daripada pengujian padanan konsep yang dilakukan. Hasil pengujian menunjukkan sekurang-kurangnya tiga konsep dari kedua-dua domain yang diekstrak adalah sepadan dengan tag <META> kata kunci yang disediakan oleh pengarang. Rajah 12 menunjukkan graf pengujian padanan konsep yang telah diplotkan berdasarkan hasil yang diperoleh dalam Jadual 1.

JADUAL 1. Keputusan pengujian padanan konsep

Konsep Pengarang	Konsep Sistem (Domain Perubatan)	Konsep Sistem (Domain Komputer)
1	0.56	0.85
2	1.18	1.43
3	1.64	1.86
4	1.96	2.00
5	2.26	2.29
6	2.48	2.43
7	2.76	2.43
8	3.12	2.57



RAJAH 12. Graf pengujian padanan konsep

Hasil ini juga selari dengan hasil pengujian yang dilakukan untuk KPSpotter [19] dan KEA [18]. 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 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 dalam talian. Kedua-dua dokumen ini adalah lebih berstruktur dan terkawal jika dibandingkan dengan dokumen web di Internet yang tiada struktur piawaian dalam perwakilan dokumen.

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

- Domain ontologi yang digunakan hanya merangkumi sebahagian kecil domain perubatan iaitu hanya 24 konsep berkaitan dengan domain jantung digunakan sebagai domain ontologi. Di samping itu juga, konsep yang disediakan pengarang kadang kala 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.
- Tag <META> kata kunci yang disediakan oleh pengarang dokumen dalam pengujian pula kadang kala 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 kadang kala bukan merupakan pilihan terbaik dalam menggambarkan kandungan maklumat dokumen. Contohnya turut dinyatakan nama pengarang sebagai tag <META> kata kunci dokumen web.
- Kaedah padanan semasa pengujian yang hanya menggunakan padanan tepat sahaja juga menyumbang hasil padanan yang rendah.

KESIMPULAN

Kajian ini memperlihatkan potensi domain ontologi dalam menyokong proses pengelasan dan pengorganisasian maklumat semantik dokumen bagi teks yang tidak berstruktur dalam laman web. Penggunaan pendekatan ontologi bukan sahaja mampu mengekstrak konsep penting dalam dokumen web malahan berupaya untuk mendapatkan maklumat kandungan semantik dokumen web.

Walaupun tesauri terkawal digunakan secara meluas dalam sistem capaian maklumat, namun ianya masih menggunakan kaedah sintaktik dalam proses padanan dengan istilah indeks dokumen (Brasethvik & Gulla 2001). Perkaitan maksud dan struktur istilah dalam tesauri terkawal tidak mewakili maksud semantik sesuatu dokumen. Sementara ontologi pula merupakan salah satu kaedah yang digunakan bagi membolehkan perkongsian maklumat antara manusia dan mesin (Bruijn 2003). Ianya boleh direalisasikan dengan semua sumber maklumat merujuk kepada suatu ontologi atau notasi yang mengandungi definisi maklumat yang mewakili sumber maklumat tersebut. Di samping itu juga, perwakilan maklumat dalam bentuk model semantik merupakan satu kaedah pengurusan maklumat yang baik (Woods 1997). Pengkayaan perwakilan semantik dokumen web diharap akan dapat menangani masalah pengindeksan dan seterusnya membantu proses pencarian dan pencapaian maklumat pada masa akan datang.

RUJUKAN

Alani, H., Kim, S., Millard, D. E., & Weal, M. J. 2003. Automatic ontology-based knowledge extraction and tailored biography generation from the web. *IEEE Intelligent System* 18(1): 14-21.

Arul Prakash Asivatham & Kranthi Kumar Ravi. 2001. Web page classification based on document structure. *IEEE National Convention.* (daalm talian) http:// citeseer.ist.psu.edu/491514.html (8 Januari 2003)

Berners-Lee, T., Hendler, J. and Lassila, O. 2001. The Semantic Web. *Scientific American* 284(5): 34-43.

Bodenreider, O. 2001. Medical Ontology Research. A report to the Board of Scientific Counselors of the Lister Hill National Center for Biomedical Communications. National Library of Medicine, (atas talian) http://lhncbc.nlm.nih.gov/cgsb/research/ umls/mor (20 November 2002)

Brasethvik, T. & Gulla, A.J. 2001. Natural language analysis for semantic document modelling. *Data And Knowledge Engineering* 38(1): 45-64.

Bruijn, J.D. 2003. Using Ontologies: Enabling knowledge sharing and reuse on the semantic web. Tesis Sarjana Digital Enterprise Research Institute (DERI), University of Innsbruck.

Desmontils, E., & Jacquin, C. 2002. Indexing a web site with terminology oriented ontology. *The Emerging Semantic Web*. Amsterdam:IOS Press: 181-198.

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.

Harmelen, F. and Fensel, D. 1999. Practical knowledge representation for the web. IJCAI-99 Workshop on Intelligent Information Integration. (dalam talian) www.cs.vu.nl/~frankh/abstracts/IJCAI99-III.html (2 Julai 2004)

Lenat, D.B. 1995. A large-scale investment in knowledge infrastructure. *Communications of the ACM* 38(1): 32-38.

Luhn, H.P. 1958. The Automatic Creation of Literature Abstract. *IBM Journal of Research and Development* 2(2): 159-165.

Rosa, M., Iocchi, L. & Nardi, D. 1998. Knowledge representation techniques for information extraction on the Web. *Proceedings of Webnet 98.* (dalam talian) www.dis.uniroma1.it/~nardi/Ricerca/papers-html/dero-iocc-nard-98. html (2 Jun 2004)

Rousseau, B. & Rousseau, R. 2002. Some idea concerning the Semantic Web. *Library and Information Service*: 39-49.

Sekine, S. 2003. Apple Pie Parser. (dalam talian) http://nlp.cs.nyu.edu/app/ (10 Januari 2003)

Smith, B. & Welty, C. 2002. Ontology: Towards a new synthesis, in: *Proceedings of the 2nd International Conference in Formal Ontology in Information Systems*: 3-9.

Smith, M.K., Welty, D. & McGuinness (peny) 2003. *OWL Web Ontology Language Guide. W3C Working Draft 31.*

Sowa, J.F. 1999. Ontological categories, in: Albertazzi (peny.), Shapes of Forms: From Getalt Psychology and Phenomenology to Ontology and Mathematics , Kluwer Academic Publishers, Dordecht : 307-340.

Song, M., Song, I.Y., & Hu, T. 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) In conjunction with the 2004 IEEE/WCI/ACM International Conference on Web Intelligence. Sept. 20-24, 2004, Beijing, China.

Snoussi, H. Magnin, L. & Nie J.Y. 2002. Toward an ontology-based web data extraction. .The AI-2002 Workshop on Business Agents and the Semantic Web (BASeWEB) held at the AI 2002 Conference (AI-2002): 26-34.

Villa, R., Wilson, R., & Crestani, F. 2003. Ontology mapping by concept similarity. *International Conference on Digital Libraries*: 666–674.

Witten I.H., Paynter G.W., Frank E., Gutwin C. & Nevill-Manning C.G. 1999. KEA: Practical automatic keyphrase extraction." *Proceedings of ACM Digital Libraries Conference*: 254-256.

Woods, W. A. 1997. Conceptual Indexing: A better way to organize knowledge. *A Sun Labs Technical Report: TR-97-61 Editor, Technical Reports.*

Zadeh, L. 2004. A note on web intelligence, world knowledge and fuzzy logic. *Data and Knowledge Engineering* 50: 291-304.

Arifah Che Al-Hadi
Jabatan Sains Komputer
Fakulti Sains dan Teknologi
Universiti Malaysia Terengganu
21030 Kuala Terengganu, Terengganu
arifah_hadi@umt.edu.my

Lailatul Qadri binti Zakaria,
Shahrul Azman Mohd Noah
Tengku Mohd Tengku Sembok
Jabatan Sains Maklumat
Fakulti Teknologi dan Sains Maklumat
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor
laila@ftsm.ukm.my
samn@ftsm.ukm.my
tmts@ftsm.ukm.my