CAPAIAN SECARA SEMANTIK DOKUMEN WEB BERASASKAN DOMAIN ONTOLOGI

ARIFAH BINTI CHE ALHADI

UNIVERSITI KEBANGSAAN MALAYSIA

CAPAIAN SECARA SEMANTIK DOKUMEN WEB BERASASKAN DOMAIN ONTOLOGI

ARIFAH BINTI CHE ALHADI

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEH IJAZAH
SARJANA TEKNOLOGI MAKLUMAT

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2005

PENGAKUAN

Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya telah saya jelaskan sumbernya.

7 Julai 2005

ARIFAH BINTI CHE ALHADI
P24712

PENGHARGAAN

Bersyukur ke hadrat Ilahi kerana dengan kurnia-Nya saya akhirnya berjaya menyiapkan kajian penyelidikan dan penulisan tesis ini. Penghargaan ini ditujukan khas buat semua insan yang banyak membantu saya dalam menyiapkan tesis sarjana ini. Terutamanya penyelia utama, Prof. Madya Dr. Shahrul Azman Mohd Noah atas segala tunjuk ajar, bantuan, bimbingan dan kritikan membina yang diberikan sepanjang penyeliaan beliau. Penglibatan yang tidak berbelah bagi dari beliau amat saya hargai. Jutaan terima kasih saya ucapkan atas segalanya. Tidak lupa juga kepada Prof. Dr. Tengku Sembok Tengku Mohd dan Puan Nazlena Mohamad Ali di atas kesudian beliau meluangkan masa memberikan pandangan membina dalam kelangsungan kajian penyelidikan ini.

Penghargaan ini turut ditujukan kepada semua kakitangan Fakulti Sains dan Teknologi, khususnya kakitangan Jabatan Sains Maklumat di atas segala kemudahan dan bantuan yang diberikan. Tidak lupa juga, kepada kakitangan Pusat Pengajian Siswazah yang banyak membantu dalam memberikan maklumat terkini pengajian.

Setinggi-tinggi penghargaan ditujukan buat Kementerian Sains, Teknologi dan Inovasi (MOSTI), di atas sokongan dan sumbangan yang diberikan di bawah perancangan Malaysia ke-lapan. Penyelidikan yang dijalankan ini adalah di bawah projek IRPA (04-02-02-0041-EA108), yang bertajuk *Development of Intelligent Tools for Managing the Semantic Information of Web Documents.*

Kepada teman-teman seperjuangan, sokongan dan bantuan yang kalian berikan telah membuahkan hasil dan banyak membantu sepanjang pengajian saya di sini. Terutamanya kepada rakan seangkatan Lailatul Qadri Zakaria yang banyak membantu dalam memberikan idea dan bimbingan.

Penghargaan ini turut ditujukan kepada staf UKM yang terlibat secara langsung dan tidak langsung dalam memudahkan perjalanan kajian penyelidikan ini. Akhirnya penghargaan buat insan tersayang, Ahmad Shafie Abdul Rahman, Muhammad Afif dan Nur Afiqah di atas sokongan dan pengorbanan dari kalian.

ABSTRAK

Capaian dan pengekstrakan maklumat semantik daripada dokumen web adalah amat penting dalam merealisasikan web semantik dan meningkatkan kualiti capaian maklumat. Pengekstrakan maklumat semantik secara manual adalah tidak praktikal dan tidak boleh diukur, sementara pengesktrakan secara automatik pula masih dalam peringkat awal untuk diimplementasikan. Oleh yang demikian, penggunaan domain pengetahuan khusus dalam bentuk ontologi dilihat sebagai salah satu alternatif penyelesaian yang boleh diambil bersesuaian untuk jangka masa singkat ini untuk mencapai dan mengekstrak kandungan maklumat semantik terutamanya bagi teks yang tidak berstruktur dalam laman web. Ontologi merupakan "perwakilan nyata bagi sesuatu domain", yang mana konsep dan hubungannya akan diisytiharkan sebagai istilah perwakilan yang membenarkan perkongsian dan penggunaan semula maklumat. Ontologi digunakan secara meluas dalam menyokong komunikasi dan perkongsian maklumat dalam konteks web semantik, namun potensinya dalam capaian semantik dokumen web masih belum diterokai sepenuhnya. Matlamat utama penyelidikan ini ialah untuk mencadangkan suatu pendekatan yang boleh diaplikasikan dalam capaian semantik ke atas dokumen yang tidak berstruktur dengan menggunakan domain ontologi. Idea utama penyelidikan ini ialah untuk menjana model semantik dokumen setiap kali pengkuerian dilakukan, yang mana pengguna dapat mencapai dan melayarinya secara semantik. Penggunaan domain ontologi spesifik adalah untuk menggambarkan kepada pengguna "perkongsian persetujuan" maksud sebenar maklumat. Dalam kajian ini, pengguna menghantar kueri berbentuk bahasa tabii. Kueri akan dianalisis dari segi struktur sintaktik dan semantik menggunakan analisis bahasa tabii dan domain ontologi dalam mendapatkan model kueri. Model semantik kueri ini kemudiannya akan dihantar kepada enjin gelintar komersial sedia ada. *Hit* (dokumen) yang diterima dari enjin gelintar akan dianalisis untuk mengekstrak calon konsep yang berpotensi menerangkan kandungan dokumen. Ayat yang mengandungi calon konsep sahaja akan dianalisis dan dibandingkan dengan domain ontologi untuk penjanaan model semantik dokumen. Domain ontologi yang digunakan dalam penyelidikan ini ialah *Medical Subject Heading* (MeSH). Setiap model semantik dokumen yang dijana akan melalui proses pengintegrasian dan hasilnya ialah penjanaan model integrasi semantik dokumen. Model integrasi semantik dokumen ini akan dipaparkan kepada pengguna yang mana ia dapat diperincikan dan dilayari secara semantik. Kajian penyelidikan ini telah menunjukkan bagaimana penggunaan domain ontologi dan teknik analisis bahasa tabii ini mampu menyokong pengkuerian semantik dokumen dengan penjanaan model semantik dokumen yang dapat diperincikan dan dilayari secara semantik. Dalam situasi ini, perincian kueri pengguna boleh dicapai secara interaktif dengan cara menyusuri model integrasi semantik dokumen.

ONTOLOGY DRIVEN TO SEMANTIC RETRIEVAL OF WEB DOCUMENTS

ABSTRACT

Accessing and extracting semantic information from Web documents is crucial for the realization of the semantic web and the provision of advance knowledge services. Manual extraction of semantic information is impractical and unscalable and fully automated tools are still at the very early stage to be implemented. Therefore, the use of specialized domain knowledge in the form of ontology is seen as one of the practical short terms solutions approach which be used to search and extract semantic information content from unstructured text on the web. Ontologies have been defined as "explicit representation(s) of domain", in which concepts and relationships between them are defined as a representational terms, enabling knowledge to be shared and reuse. While ontologies have been widely used to support communication and information sharing within the context of Semantic Web development, their potential use to support the task of semantic querying of web documents is still not thoroughly explored. The main aim of this research is therefore to propose an approach of which the unstructured nature of web documents can be semantically queried by using domain ontology. The main inherent idea is to generate a semantic document model for each search session which can be semantically searched and browsed by the user. The use of domain specific ontology is to reflect the users' "shared agreement" on the meaning of information. In this approach, users will submit their query in the form of natural language. This query will be syntactically and semantically analysed using natural language analysis (NLA) technique and domain ontology to form a query model. The query model will then be submitted to existing commercial search engine. Hits received from the search engine will be analysed to extract candidate concepts that are potential to describe the document content. Using this approach only those sentences that are related to the candidate concepts are analysed and compared with the domain ontology to construct the semantic document model. In this research, the Medical Subject Heading (MeSH) domain ontology has been used. Each generated semantic model will then undergo an integration process, which results in the creation of an integrated semantic document model. The integrated semantic document will be presented to users as the query of which can be semantically refined and browsed. This research has demonstrated how a domain ontology combined with NLA technique can be exploited to support the task of semantic document querying by constructing a semantic document model which can be further refined and browsed. In this sense, refinement of the user's query can be interactively achieved by browsing the generated integrated semantic document model.