PENGEKSTRAKAN DATA BERASASKAN PENDEKATAN ONTOLOGI : KES DATA JUJUKAN HIDROLOGI

AHMAD GHADAFFI ABD HAMID

UNIVERSITI TEKNOLOGI MALAYSIA

PENGEKSTRAKAN DATA BERASASKAN PENDEKATAN ONTOLOGI : KES DATA JUJUKAN HIDROLOGI

AHMAD GHADAFFI BIN ABD HAMID

Tesis ini dikemukakan
sebagai memenuhi syarat penganugerahan
ijazah Sarjana Sains (Sains Komputer)

Fakulti Sains Komputer dan Sistem Maklumat
Universiti Teknologi Malaysia

DISEMBER 2005

Buat Isteri dan Anak Tercinta,
AyahandaBonda di Kuala Terengganu dan Ipoh,
Keluarga tersayang
Terima kasih atas segala dorongan dan pengorbanan…

PENGHARGAAN

Syukur ke hadrat Ilahi kerana dengan izinNya tesis ini dapat disiapkan. Setinggi-tinggi penghargaan kepada penyelia tesis, Prof. Madya Dr. Harihodin Selamat, Prof. Madya Daut bin Daman dan En. Mohd Shafry bin Mohd Rahim atas bimbingan dan penyeliaan yang diberi sepanjang tempoh penyediaan tesis. Saya juga terhutang budi diatas kesudian mereka membiayai pengajian sarjana ini.

Penghargaan yang tidak terhingga juga ditujukan buat isteri tercinta Puteri Suhaiza Sulaiman yang banyak memberi pandangan dan kritikan. Tanpa sokongan beliau, tesis ini tidak akan sama seperti yang dibentangkan disini.

Sekian, terima kasih.

ABSTRAK

Pengekstrakan maklumat merupakan satu proses yang mengekstrak maklumat daripada sumber sistem sedia ada dan menyimpannya ke dalam pangkalan data. Penyelidikan terdahulu tertumpu kepada pengekstrakan maklumat data HTML menggunakan pendekatan *wrapper*. Kelemahan pendekatan ini adalah dari segi ketahanan di mana *wrapper* gagal berfungsi dengan baik jika terdapat perubahan pada struktur fail yang ingin di ekstrak. Pengekstrakan maklumat berasaskan ontologi merupakan penyelesaian alternatif kepada masalah ketahanan. Di dalam penyelidikan ini, pengekstrakan maklumat berasaskan ontologi menggunakan data hidrologi dari Jabatan Pengairan dan Saliran (JPS) sebagai kajian kes. Pengekstrakan maklumat ontologi bagi domain hidrologi dikenali sebagai 'EkstrakPro' terbahagi kepada tiga proses utama; iaitu proses penghuraian ontologi, proses pengecam jujukan dan kata kunci serta proses pemetaan data. 'EkstrakPro' menggunakan dua input; data hidrologi dan ontologi pengekstrakan. Ciri penting 'EkstrakPro' adalah ontologi pengekstrakan, di mana unit objek diperkenalkan bagi memudahkan selenggara ontologi. Algoritma pengecam jujukan menyelesaikan isu penggunaan masa dalam mengekstrak data berjujukan. Lima jenis data hidrologi digunakan di dalam eksperimen. Data-data ini dibahagikan kepada tiga kategori; (i) Data asal daripada mesin bacaan, (ii) data yang diubahsuai dan (iii) perbezaan saiz data. Berdasarkan kategori tersebut, ketahanan pengekstrakan maklumat dan masa yang digunakan dapat diukur menggunakan rumusan ketepatan dan notasi-O. Keputusan menunjukkan prototaip 'EkstrakPro' boleh mengekstrak data hidrologi dengan struktur yang berbeza dengan tepat dan menggunakan hanya satu algoritma. Algoritma pengecam jujukan boleh juga mengurangkan masa yang diperlukan oleh pengekstrakan maklumat. Hasil penyelidikan ini membuktikan masalah pengekstrakan maklumat dapat diselesaikan dengan pendekatan ontologi.

ABSTRACT

Information Extraction is a process that extracts information from existing system source and stores into a database. Previous researchers had focus on information extraction for HTML data using wrapper approach. The drawback from this approach is resiliency where wrapper fails to function when the file of interest's structure changes. Ontology based information extraction is an alternative solution for this problem. In this research, ontology based information extraction used hydrological data from Jabatan Pengairan dan Saliran (JPS) as the case study. Ontology based information extraction for hydrology domain or also known as 'EkstrakPro' is divided into three main processes; which are ontology parser process, keyword and sequences recognition process, and a data mapping process. 'EkstrakPro' used two inputs; the hydrology data and ontology extraction. An important feature in 'EkstrakPro' is that ontology extraction, where unit object is introduced to simplify the ontology maintenance. The sequential recognition algorithm is to solve the time consuming issues for extracting sequential data. Five types of hydrological data are used in the experiment. These data are divided into three categories; (i) original data taken from gauging machine, (ii) the altered data and (iii) the different sizes of data. Based on these categories, the information extraction resiliency and time taken have been measured using a precise equation and O-notation. The results show that prototype 'EkstrakPro' can extract different structure hydrology data correctly by using only one algorithm. Using sequential recognition algorithm can also further reduce the time required for extraction of information. The result of the research proves that information extraction can be solved using ontology approach.

KANDUNGAN

BAB	TAJUK	MUKA SURAT
1	**PENGENALAN**	
1.1	Pendahuluan	1
1.2	Latar Belakang Masalah	2
1.3	Kajian Kes	4
1.4	Motivasi Kajian Kes	5
1.5	Pernyataan Masalah Penyelidikan	5
1.6	Matlamat Penyelidikan	6
1.7	Objektif Penyelidikan	6
1.8	Skop Penyelidikan	6
1.9	Sumbangan Tesis	7
1.10	Struktur Tesis	8
2	**KAJIAN LITERASI**	
2.1	Pendahuluan	9
2.2	Pengekstrakan Maklumat (IE)	9
	- Bahasa Pembangunan Wrapper	10
	- Pendekatan HTML	10
	- Pendekatan Induksi	10
	- Pendekatan Model	11
	- Pendekatan NPL	11

- Pendekatan Ontologi 11
2.3 Pengekstrakan Berasaskan Ontologi 13
2.4 Ontologi Pengekstrakan 16
2.5 Kajian Kes ke atas Data Hidrologi JPS 18
2.5.1 SRM 18
2.5.2 MIT 20
2.5.3 CSV 21
2.6 Kesimpulan 21

3 METODOLOGI PENYELIDIKAN

3.1 Pendahuluan 22
3.2 Ontologi Pengekstrakan 24
3.2.1 Penggunaan OSM 24
3.2.2 Unit Objek 26
3.2.2.1 Stesen_Id 28
3.2.2.2 Nama_stesen 28
3.2.2.3 Jenis_cerapan 28
3.2.2.4 Tarikh_cerapan 29
3.2.2.5 Masa_cerapan 29
3.2.2.6 Nilai_cerapan 30
3.3 Proses Penghuraian Ontologi 30
3.4 Proses Pengecam Jujukan 32
3.5 Proses Pemetaan 36
3.6 Pengujian 37
3.7 Kesimpulan 37

4 IMPLEMENTASI

4.1 Pendahuluan 39
4.2 Spesifikasi Sistem 39
4.3 Antara Muka Sistem 40
4.4 Implementasi Proses Penghurai Ontologi 42
4.5 Implementasi Proses Pengecam Jujukan dan Katakunci 44
4.6 Implementasi Proses Pemetaan Data 45
4.7 Kesimpulan 45

5 PENGUJIAN

5.1 Pendahuluan 46
5.2 Penyediaan Data Ujian 46
5.3 Ujian Ketahanan Pengekstrakan Data 47
5.4 Ujian Masa Pengekstrakan Data 49
5.5 Kesimpulan 52

6 KESIMPULAN

6.1 Pendahuluan 53
6.2 Rumusan Keseluruhan Penyelidikan 53
6.3 Kebaikan dan Kelemahan Kajian 55
6.4 Penambahbaikan 56
6.5 Penutup 56

RUJUKAN 57
Lampiran A - F 62 - 84

SENARAI JADUAL

NO JADUAL	TAJUK	MUKA SURAT
3.1	Ringkasan metodologi penyelidikan	38
5.1	Peratus ketepatan bagi algoritma *MHIS Dataload* dan algoritma EkstrakPro	47

SENARAI RAJAH

NO RAJAH	TAJUK	MUKA SURAT
1.1	Struktur Tesis	8
2.1	Rangka Kerja Pengekstrakan Maklumat Berasaskan Ontologi	13
2.2	Contoh Dokumen Tidak Berstruktur	14
2.3	Contoh keratan format SRM	19
2.4	Penyusunan format SRM	20
2.5	Contoh keratan format MIT	20
2.6	Contoh Keratan format CSV	21
3.1	Reka Bentuk Embley et al.(1998) Dengan Penambahan Proses Pengecam Jujukan	23
3.2	Ontologi data hidrologi JPS secara grafikal	25
3.3	Ontologi data hidrologi JPS secara teks	26
3.4	Sintek Rangka UO	27

3.5	Contoh Stesen_Id daripada data hidrologi JPS	28
3.6	Contoh Tarikh_cerapan daripada data hidrologi JPS	29
3.7	Contoh Masa_cerapan daripada data hidrologi JPS	30
3.8	Skema pangkalan data daripada ontologi pengekstrakan	31
3.9	Algoritma EkstrakPro	32
3.10	Corak jujukan data hidrologi JPS	33
3.11	Notasi algoritma pengecaman jujukan	34
3.12	Algoritma pengecaman jujukan	35
3.13	Algoritma EkstrakPro dengan Algoritma jujukan	36
4.1	Antara muka *EkstrakPro*	39
4.2	Reka Bentuk Sistem dan Antara Muka Sistem EkstrakPro	40
4.3	Input Ontologi Pengekstrakan bagi Tarikh Cerapan	41
4.4	Keratan Atur cara Penghuraian Ontologi	42
4.5	Contoh Skema Pangkalan Data	43
4.6	Keratan Aturcara Pengekstrakan Katakunci	43
4.7	Keratan Pernyataan *Insert*	44

5.1	Peratus ketepatan pengekstrakan data terhadap jenis data	48
5.2	Perbandingan masa pengekstrakan dengan algoritma pengecam jujukan dan tanpa algoritma pengecam jujukan	50

SENARAI SINGKATAN

AI	-	*Artificial Intelligent*
BYU	-	*Brigham Young University*
CSV	-	*Comma Separated Variable*
IE	-	*Information Extraction*
JPS	-	Jabatan Pengairan dan Saliran
MHIS	-	*Malaysian Hydrology Information System*
MIT	-	*Molecule Information Table*
NPL	-	*Natural Language Processing*
SRM	-	*Single Robust Model*
UO	-	Unit Objek

SENARAI LAMPIRAN

NO LAMPIRAN	TAJUK	MUKA SURAT
A	Contoh rangka unit objek bagi stesen ID	62
B	Contoh rangka unit objek bagi tarikh cerapan	64
C	Contoh rangka unit objek bagi masa cerapan	67
D	Contoh keratan data hidrologi kategori pertama	70
E	Contoh keratan data hidrologi kategori kedua	72
F	Contoh keratan data hidrologi kategori ketiga	82

BAB 1

PENGENALAN

1.1 Pendahuluan

Bidang *Information Extraction* (IE) adalah satu bidang yang melakukan proses pengekstrakan maklumat daripada data digital. Youn (1992) mendefinisikan pengekstrakan maklumat sebagai satu proses untuk mengekstrak maklumat daripada sumber sistem sedia ada dan seterusnya menyimpannya ke dalam satu fail. Manakala Xiaoying dan Mengjie (2004) mendefinisikan IE sebagai satu proses yang mengambil fail teks sebagai input dan menghasilkan data mengikut format yang diperlukan. Data ini mungkin dipaparkan kepada pengguna, disimpan di dalam pangkalan data atau *spreadsheet* bagi kegunaan analisis.

Di antara kepentingan IE yang dikenal pasti adalah membantu enjin pencarian dokumen daripada halaman web. Teknik pengekstrakan diperlukan dalam mencari maklumat yang tepat daripada satu atau lebih dokumen web. Selain itu IE diperlukan dalam proses pemindahan data daripada sistem asal ke sistem yang baru. Situasi ini sering berlaku apabila pengguna bertukar sistem komputer. Data daripada sistem asal akan di ekstrak dan diubah format yang sesuai dengan sistem yang baru.

Terdapat beberapa pendekatan IE termasuklah bahasa pembangunan *wrapper,* penggunaan struktur data, *Natural Language Processing* (NLP), permodelan dan ontologi. Tumpuan kebanyakan penyelidik adalah meningkatkan ketepatan *wrapper* di samping mengurangkan penglibatan pengguna dalam proses pengekstrakan iaitu secara automatik. Kelemahan utama sistem IE yang

menggunakan pendekatan *wrapper* adalah ia hanya dapat mengekstrak maklumat daripada data dalam berformat yang terhad dan tertentu sahaja.

Sementara itu, terdapat sekumpulan penyelidik daripada Universiti Brigham Young sedang berusaha meningkatkan penggunaan konsep skema yang lebih umum bagi meningkatkan ketepatan IE. Kumpulan ini mula memperkenalkan pendekatan ontologi di dalam IE (Embley et al., 1998). Ontologi adalah spesifikasi dalam membentuk suatu konsep (Gruber, 1993). Dari sudut bidang falsafah, ontologi merujuk kepada suatu kewujudan. Di dalam konsep perkongsian pengetahuan (*knowledge sharing*) aplikasi kepintaran buatan (AI), ontologi adalah penerangan mengenai konsep dan hubungan yang wujud bagi satu agen. Kelebihan utama IE berasaskan ontologi adalah mempunyai ketahanan pengekstrakan maklumat. Menyedari kelebihan ini, bidang IE berasaskan ontologi akan menjadi fokus penyelidikan ini.

1.2 Latar Belakang Masalah

Penggunaan data digital telah berkembang pesat beberapa tahun kebelakangan ini. Ini kerana dorongan penggunaan *world web wide* (www) yang semakin meningkat. IE digunakan bagi mengekstrak maklumat daripada fail HTML. Pendekatan seperti bahasa *wrapper* (Crescenzi et al., 2001; Hammer et al., 1997; Arocena dan Mendelzon, 1998), NLP (Calif dan Mooney, 1999; Freitag, 2000;Sonderlan,1999) dan permodelan (Adelberg, 1998) diperkenalkan bagi mengekstrak maklumat yang diperlukan pengguna. Walaupun kebanyakan penyelidik melaporkan kejayaan hasil daripada pengujian yang dilakukan, namun pendekatan ini masih mempunyai masalah ketahanan. Kelemahan dari segi ketahanan bermakna sebuah *wrapper* akan gagal berfungsi dengan baik sekiranya terdapat perubahan pada struktur fail yang ingin di ekstrak.

IE berasaskan ontologi adalah penyelesaian kepada masalah ketahanan. Pengekstrakan maklumat ontologi adalah model konsepsi yang menerangkan aplikasi

dunia sebenar dengan terperinci. Ciri penting pendekatan ini adalah ontologi pengekstrakan yang dihasilkan daripada data dalam sesebuah bidang tanpa bergantung kepada struktur fail input.

Oleh sebab kebanyakan IE berasaskan ontologi hanya tertumpu kepada fail HTML, timbul persoalan, apakah pendekatan ini boleh digunakan ke atas dokumen lain selain fail HTML? Dalam penyelidikan kali, kajian akan dilaksanakan ke atas IE berasaskan ontologi dengan menggunakan fail teks. Ini kerana fail teks mengandungi sedikit penunjuk untuk mengenal pasti struktur berbanding dengan fail HTML. Fail HTML mempunyai penunjuk-penunjuk yang membezakan struktur antara permulaan <head>, tajuk <title>, kandungan <body> dan sebagainya. Sementara itu tidak semua elemen di dalam fail teks dipisahkan dengan tanda atau tag HTML. Maka proses IE daripada fail teks adalah lebih sukar daripada fail HTML (Adelberg, 1998).

Menyedari kekurangan penyelidikan ke atas IE berasaskan ontologi bagi data selain HTML, penyelidikan ini telah memilih untuk mengkaji keberkesanan IE berasaskan ontologi dalam mengekstrak data hidrologi. Satu kajian kes dilakukan ke atas *Malaysian Hidrology Information System* (MHIS) dari Jabatan Pengairan dan Saliran (JPS), yang mana sebelum ini menggunakan pendekatan pengekstrakan data secara tradisional. Penerangan dan kelemahan MHIS akan dibincangkan pada Bahagian Kajian Kes.

1.3 Kajian Kes

MHIS di Jabatan Pengairan dan Saliran (JPS) telah dibangunkan dengan usaha sama Universiti Teknologi Malaysia (UTM) dan *Water Institute*, UK. MHIS digunakan untuk menyimpan dan manipulasi maklumat hidrologi yang terdiri daripada beberapa modul antaranya adalah perisian *MHIS Dataload*. Modul ini menyediakan kemudahan untuk memindahkan data hidrologi ke dalam sistem pangkalan data MHIS (Jabatan Pengairan dan Saliran, 2001a).

MHIS Dataload terdiri daripada beberapa algoritma yang dibangunkan khas bagi data taburan hujan, penyejatan, aras air sungai, enapan terapung dan kualiti air. Algoritma pengekstrakan data telah ditulis di dalam atur cara secara tetap (*hardcoded*) bagi setiap jenis data-data di atas. Proses penyenggaraan perisian ini memerlukan banyak usaha dan masa. Berikut adalah beberapa kelemahan *MHIS Dataload* yang telah dikenal pasti :

1. Algoritma mengekstrak data tidak dinamik. Maka algoritma perlu dikemas kini apabila perubahan struktur atau format data berlaku. Perisian perlu dikemaskinikan setiap kali berlaku perubahan struktur data.

2. Satu algoritma digunakan bagi satu jenis data hidrologi. Maka apabila satu jenis data hidrologi baru digunakan, ia memerlukan satu algoritma pengekstrakan yang baru.

3. Algoritma bergantung kepada struktur dan format data. Data yang dihasilkan oleh manusia selalunya mempunyai banyak ralat atau kesilapan. Data yang akan di ekstrak perlu dibersihkan daripada kesilapan dan ralat.

Berdasarkan kelemahan-kelemahan di atas, persoalan yang dikaji adalah apakah IE berasaskan ontologi sesuai untuk data hidrologi dan sekali gus dapat mengatasi kelemahan-kelemahan yang dihadapi oleh MHIS *Dataload* ?

1.4 Motivasi Kajian Kes

Penyelesaian yang dihasilkan di dalam penyelidikan ini akan dapat membantu dalam mempertingkatkan kecekapan dan ketepatan kerja-kerja pemindahan data hidrologi di dalam bentuk teks ke dalam pangkalan data *MHIS* di JPS.

1.5 Pernyataan Masalah Penyelidikan

Tujuan penyelidikan ini adalah untuk mengkaji IE berasaskan ontologi dengan menggunakan fail teks hidrologi JPS. Dengan implementasi ontologi pengekstrakan ke atas bidang data hidrologi, perkara berikut perlu diperjelaskan.

1. Bagaimana menghasilkan ontologi pengekstrakan bagi mencapai matlamat penyelidikan?
2. Bagaimana menyatakan dengan cara teratur bagi setiap kata kunci, prosa bidang data hidrologi?
3. Bagaimana maklumat diasingkan daripada sumber data berdasarkan kata kunci di dalam ontologi?
4. Bagaimana menentukan keberkesanan IE berasaskan ontologi mengekstrak maklumat daripada fail teks hidrologi.
5. Apakah pembaikan yang boleh dilakukan ke atas IE berasaskan ontologi dalam mengekstrak fail teks hidrologi.

1.6 Matlamat Penyelidikan

Mengkaji keberkesanan IE berasaskan ontologi dalam mengekstrak maklumat daripada fail teks bidang hidrologi.

1.7 Objektif Penyelidikan

Objektif penyelidikan adalah seperti berikut :

1. Membina ontologi pengekstrakan bagi menterjemahkan kata kunci dan hubungan kata kunci fail teks hidrologi.
2. Membina algoritma pengecam jujukan bagi mengurangkan masa pengekstrakan.
3. Melakukan pengujian pengekstrakan maklumat daripada fail teks hidrologi.

1.8 Skop Penyelidikan

1. Fail yang digunakan adalah fail teks berjujukan, yang mana bentuk jujukan adalah konsisten. Fail input yang digunakan adalah data hidrologi daripada JPS, yang mana ia berada di dalam bentuk berjujukan.
2. Struktur pangkalan data yang digunakan berdasarkan skema yang dijana daripada ontologi pengekstrakan.
3. Ontologi pengekstrakan dihasilkan secara manual bagi menghasilkan ekspresi yang lengkap agar matlamat penyelidikan dicapai.

1.9 Sumbangan Ilmiah

Sumbangan akhir penyelidikan adalah seperti berikut :

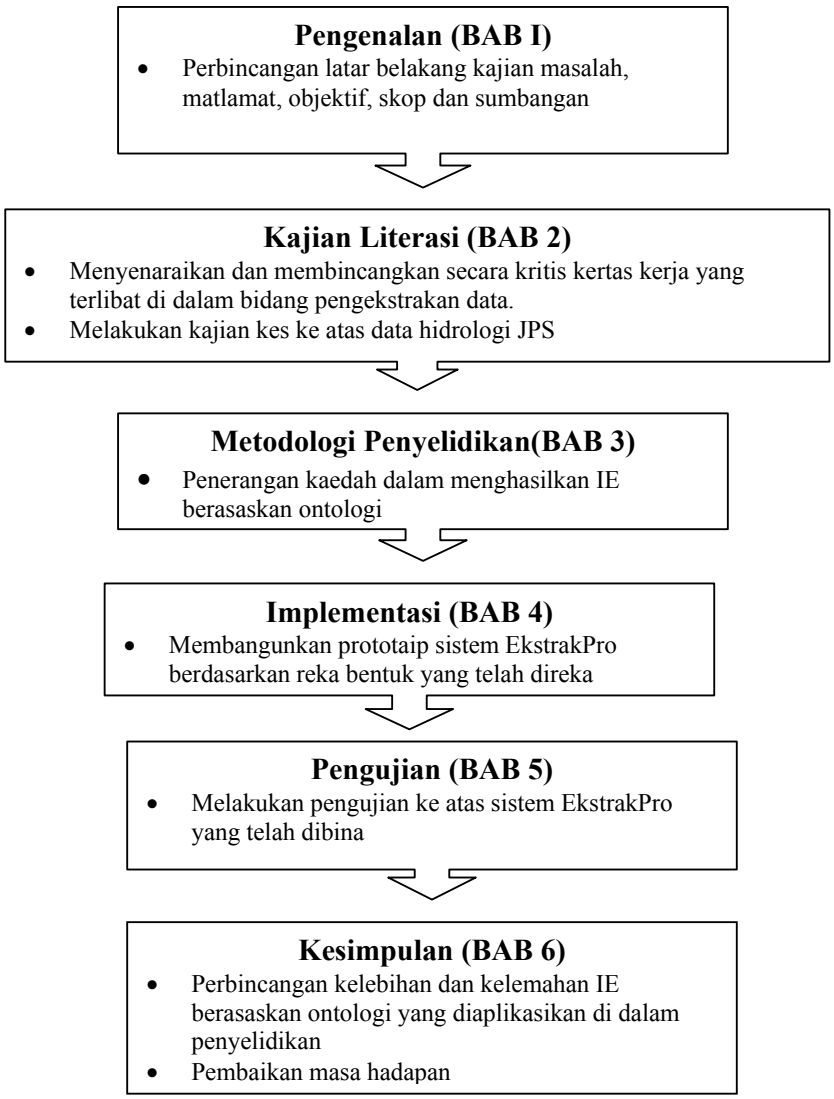
1. Di dalam penyelidikan ini, IE berasaskan ontologi telah digunakan ke atas bidang data hidrologi. Kajian menunjukkan bahawa IE berasaskan ontologi dapat digunakan ke atas fail teks berjujukan.

2. Unit Objek (UO) diperkenalkan bagi menyatakan corak kata kunci. UO adalah kaedah menghasilkan kata kunci secara sistematik. Penggunaan UO dapat mengurangkan kesilapan di dalam menghasilkan kata kunci.

3. Penghasilan algoritma jujukan dalam meningkatkan kepantasan masa proses pengekstrakan bagi data berjujukan. Algoritma pengecam jujukan berfungsi sebagai pembaca bentuk jujukan maklumat. Jika bentuk jujukan telah dikenal pasti, maklumat akan di ekstrak tanpa membandingkan kata kunci dan fail teks. Dengan ini dapat masa proses pengekstrakan dapat dipercepatkan.

1.10 Struktur Tesis

Tesis ini secara keseluruhannya terbahagi kepada 6 bab seperti ditunjukkan di dalam Rajah 1.1.



Rajah 1.1 : Struktur Tesis

RUJUKAN

Abiteboul, S. (1997). *Querying semi-structured data.* In Database Theory, 6th International Conference. January 8-10. Greece. 1-18.

Adelberg, B.(1998). *NoDoSE: A Tool for Semi-Automatically Extracting Structured and Semi-Structured Data from Text Documents*. SIGMOD Record 21(2): 283-294.

Arocena, G.O. and Mendelzon, A, O. (1998). *WebOQL: Restructuring Documents, Databases and Webs*. In Proceedings of the 14th IEEE International Conference on Data Engineering. Florida. 24-33.

Azmi Jafri (2002). *Arahan Kerja: Pemungutan data Hidrologi*. Technical Report. UPMBH-PK(AK)-01

Baumgartner, R., Sergio, F., Georg G. (2001). *Visual Web information extraction with Lixto.* In Proceedings of the 26th International Conference on Very Large Database Systems.119-128.

Brian, S., Rajeev M., Lawrence P., Terry W. (1998). *What can you do with a Web in your pocket?* Data Engineering Bulletin 21(2): 27-47.

Buneman, P. (1997). *Semi structure Data.* In Proceedings of the 16th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. Arizona. 117-121.

Califf, M. E. and Mooney, R. J.(1999). *Relational learning of Pattern-Match Rules for Information Extraction.* In Proceedings of 16th National Conference on Artificial Intelligence and 11th Conference on Innovative Applications of Artificial Intelligence. . 328-334.

Cocchiarella, N.B. (1991). *Formal Ontology*. Handbook of Metaphysics and Ontology. Munich.

Crescenzi, V. and Mecca, G. (1998) Grammars Have Exceptions. *Information System* 23(8). 539-565.

Crescenzi, V., Giansalvatore, M., Paolo, M. (2001). *RoadRunner: Towards Automatic Data Extraction from Large Web Sites*. In Proceedings of the 26th International Conference. Italy. 109-118.

Embley D. W., Barry D. K, Scott N. W. (1992). *Object oriented Systems Analysis: A Model-Driven Approach*. Prentice Hall, Englewood, New Jersey.

Embley D. W., Douglas M. C., Stephen W. L., Randy D. S.(1998). *Ontology-based extraction and structuring of information from data-rich unstructured documents*. In Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM'98), pages 52–59.

Florescu, D., Alon Levy, Alberto Mendelzon (1998). *Database techniques for the World-Wide Web: A Survey.* SIGMOD Record: 27(3). 59-74.

Freitag, D. (2000). *Machine learning for information in Informal Domains*. Machine Learning 2/3. 162-2002.

Golgher, P. B., Altigran S. da Silva., Alberto H. F. Laender, Berthier Ribeiro-Neto (2001). *Bootstrapping for Example-Based Data Extraction*. In Proceedingof the 10th ACM International Conference on Information and Knowledge Management. Georgia.

Guarino, N. (1998). Some Ontological Principles for Designing upper level lexical resources. Proceeding of the first Iternational conference on lexical resources and evaluation. Granada.

Hammer, J., H. Garcia, M., S. N., R. Yerneni, M. M. Breunig, and V. V. (1997). *Template-Based Wrappers in the TSMMIS System*. SIGMOD Record: 26(2). 532-535.

Hammer, J. (1997). *The TSIMMIS Experience.* In Proceedings of the First East-European Symposium on Advances in Databases and Information Systems. Rusia. 1-8.

Hsu, C. n. and Dung, M. T. (1998). *Generating Finite-State Transducers for Semi-Structured Data Extraction from the web*. Information Systems 23(8). 521-538.

Huck, G., Peter F., Karl A., Erich N. (1998). *jedi: Extracting and Synthesizing Information from the web*. In Proceeding of the 3rd IFCIS International Conference on cooperative Information Systems. New York. 32-43.

Hwang, C.H. (1999). Incompletely and Imprecisely Speaking: Using Dynamic Ontologies for representing and retrieving information. Proceeding of 6th international workshop on knowledge representation meets databases, KRDB'99.Sweden.

Jabatan Pengairan dan Saliran Malaysia. (2001a). *Malaysian Hydrological Information System, Final Report. Vol 1*. Technical Report

Jabatan Pengairan dan Saliran Malaysia. (2001b). *Malaysian Hydrological Information System, Final Report. Vol 2.* Technical Report

Jabatan Pengairan dan Saliran Malaysia. (2001c). *Malaysian Hydrological Information System, Final Report. Vol 3.* Technical Report

Jabatan Pengairan dan Saliran Malaysia. (2001d). *Malaysian Hydrological Information System, Final Report. Vol 4*. Technical Report

Kushmerick, N. (2000). Wrapper Induction: Efficiency and Expressiveness. *Artificial Intelligence Journal* 118(1/2): 15-68.

Laender, A.H.F, Berthier, A, Ribeiro-Neto, da Silva, Altigran Soares (2001). DEByE – data Extraction by Example. Data and Knowledge Engineering.

Laender, A. H. F (2000). *Representing Wed Data as Complex Objects.* In Electronic Commerce and Web technologies. Berlin. 216-228.

Lopez, F. M.(1999). *Overview of Methodologies for Building Ontologies.* Proceeding of IJCAI-99 Workshop on Ontologies and Problem Solving Method. Stockholm.

Mecca, G., Atzeni, P., Masci, A., Merialdo, P., Sindoni, G (1998). The ARANEUS Web-base management System. *SIGMOD*
Record: 27(2). 544-546.

Muslea, I. (1999). *Extraction Pattern for Information Extraction tasks: A survey*. In Proceeding of the AAAI-99 Workshop on Machine Learning for Information Extraction. Florida. 1-6.

Muslea, I., Steven Minton, Craig A. Knoblock (2001). *Hierarchical Wrapper Induction for Semi-structured Information Sources*. Autonomous Agents and Multi-Agent. 4(1/2).

Ribeiro-Neto, B. A. (1999). *Extracting Semi-Structured Data Through Examples*. In Proceeding of the 8th ACM International Conference on Information and Knowledge management. Missouri. 94-101.

Reich, J.R.(1999). Ontological Design Patterns for The Integration of Molecular Biological Information. Proceeding of German Conference on Bioinformatic GCB'99.October Hannover. 156-166.

Sahuguet, A. and Azavant F. (2001). *Building Intelligent Web Application using Lightweight wrappers.* Data and Knowledge Engineering 36(3): 283-316.

Soderlan, S. (1999). *Learning Information Extraction Rules for Semi-Structure and Free Text*. Machine Learning 24 (1/3): 233-272.
Uschold, M., Gruninger, M.(1996). Ontologies: Principles, Method and Application. Knowledge Engineering Review 11(2). 93-155.

Xiaoying, G. and Mengjie, Z. (2004). A Knowledge Learning Approach toInformation Extraction From Multiple Text Based Web Site. *International Jurnal On Artificial Intelligence Tools*. Vol.13, No.3:721-738.

Youn, C.(1992) .Data Migration. *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics,* Chicago, Illinois, USA, Volume 2, pages 1255-1258.