ONTOLOGY DRIVEN
WEB DATA EXTRACTION

Hazlinda Bt Ghazali

Master of Science Intelligent System (IS)

ONTOLOGY DRIVEN WEB DATA EXTRACTION

A thesis submitted to the Faculty of Information Technology in partial

fulfillment of the requirements for the degree

Master of Science (Intelligent System)

Universiti Utara Malaysia

By

Hazlinda binti Ghazali

JABATAN HAL EHWAL AKADEMIK
(Department of Academic Affairs)
Universiti Utara Malaysia

PERAKUAN KERTAS KERJA PROJEK
(Certificate of Project Paper)

Saya, yang bertandatangan, memperakukan bahawa
(I, the undersigned, certify that)

HAZLINDA GHAZALI

calon untuk Ijazah
(candidate for the degree of **MSc. (Intelligent System)**

telah mengemukakan kertas projek yang bertajuk
(has presented his/ her project paper of the following title)

ONTOLOGY DRIVEN WEB DATA EXTRACTION

seperti yang tercatat di muka surat tajuk dan kulit kertas projek
(as it appears on the title page and front cover of project paper)

bahawa kertas projek tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan.
(that the project paper acceptable in form and content, and that a satisfactory knowledge of the filed is covered by the project paper).

Nama Penyelia Utama
(Name of Main Supervisor) : **DR. NORITA MD NORWAWI**

Tandatangan
(Signature) : Tarikh *(Date)*: 1/7/04

Nama Penyelia Kedua
(Name of 2nd Supervisor) : **PUAN NORLIZA KATUK**

Tandatangan
(Signature) : Tarikh *(Date)*: 1/7/04

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence by the Dean of the Graduate School. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to

Dean of Faculty of Information Technology
Universiti Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman

ABSTRAK

(Bahasa Melayu)

Dewasa ini, pengekstrakan data dari halaman web menjadi semakin popular dan diaplikasikan secara meluas dalam pelbagai bidang. Objektif utama kajian ini ádalah untuk membangunkan satu teknik pengekstrakan data dari halaman web pengumuman persidangan IEEE dalam mengenalpasti tarikh-tarikh penting dalam satu persidangan. . Pada masa kini, pelbagai masalah timbul berikutan ketidak seragaman dan format bebas yang digunakan dalam dokumen web. Disamping itu, pelbagai terma yang sedia ada mempunyai maksud yang sama. Dalam kajian ini, maklumat daripada halaman web diekstrak dan distruktur melalui penggunaan ontologi dan data yang telah diekstrak disimpan di dalam dokumen XML. Teknik ini dibangunkan menggunakan bahasa pengaturcaraan Cold Fusion 4.5.

ABSTRACT

(English)

Data extraction from web document is becoming more popular and widely used for many tasks. The objective of this study is to develop web data extraction technique from IEEE Conference announcement website to search for important dates related to conference. At present, problems arise due to non-standardized and free format web document. Besides that, multiple terms can have same meaning. In this study, information from web pages were extracted and structured from the websites by using ontology and used XML document to store data. The web data extraction technique is developed using Cold Fusion 4.5 web programming language.

ACKNOWLEDGEMENTS

Praise be to Allah S.W.T, whose blessing and guidance have helped me through entire project works. Peace be upon our Prophet Muhammad S.A.W, who has given light to mankind. My sincere appreciation to my parents (Ghazali Hj Talib and Hamimah Mat Damin) and family for their patience, prayers and understanding over the entire period of my study.

Special thanks to my supervisors, Dr. Norita Md Norwawi and Mrs Norliza Katuk for their continuous assistant, support and supervision during the project development and in the preparation of this thesis. To Mr Azizi Ab Aziz thanks for the guidance and advice that you have given.

Equally deserving of this recognition are brilliant graduate Intelligent System batch June 2003 and lecturers who have supported me towards the completion of this project as well as my friends Hasnatul, Kamal, Muziah, Mazura, Suziane, Kak Lia, AI seniors and AI juniors for their encouragement.

Finally, to my beloved best friends, Irham, Faiz, Nat, Nona, Rudi, Olin and all the individuals who share my good and bad days.

CONTENTS

CHAPTER THREE : METHODOLOGY

CHAPTER FOUR : FINDING AND DISCUSSION

CHAPTER FIVE : CONCLUSIONS AND RECOMMENDATIONS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Acronym	**Meaning**
HTML	Hypertext Markup Language
IIS	Internet Information Services
NLP	Natural Language Processing
PWS	Personal Web Server
XML	Extensible Markup Language

CHAPTER ONE

INTRODUCTION

The World Wide Web (WWW) is a vast and rapidly growing source of information and plays the most important sources for data. It becomes one of the important media that can be used to get a lot of information. The data available on web is easy to understand by human but it is difficult to understand by machine. Extracting relevant data that is necessary by human is not a simple task. Web data extraction is a technique to select a specific portion of information from web documents and stored into databases. Most of this information is in the form of unstructured text, which makes the information difficult to query.

Extracting structured data from the web pages is clearly very useful, since it enables us to pose complex queries over the data. Structured data extraction has also been recognized as an important sub-problem in information integration systems (Haas *et al.*, 1997; Molina *et al.*, 1997; Ullman, 1997; Levy *et al.*, 1996) which integrate the data present in different web-sites. However it is not an easy task, since web documents do not have consistent format or structure. They are free format text document. Although they are structured, it is not easy to find the structures of the data. Therefore, there has

REFERENCES

AgentCities.NET. IST project IST-2000-28384 Agentcities. Retrieved 20 April 2004 from http://www.agenticities.net/

Alani, H., Kim, S., Millard, D., Weal, M., Hall, W., Lewis, P., and Shadbot, N. (2003). Automatic ontology based knowledge extraction from web documents. *IEEE Intelligent Systems*, 18, (1), pp. 14–21.

Aldea, A., Banares-Alcántara,R., Bocio, J., Gramajo, J., Isern, D., Kokossis, A., Jimenez, L., Moreno, A.,Riano, D. (2003) . An Ontology-Based Knowledge Management Platform. IJCAI Workshop on *Information Integration on the Web.* Acapulco, México, pp. 177-182.

Apers, P. M. G. (1994). Identifying internet-related database research. *In Proceedings of the 2nd International East-West Database Workshop.* Klagenfurt: Springer-Verlag, pp. 183–193.

Arasu, A. and Molina, H. G. (2003). Extracting Structured Data from Web Pages. *In Proceedings of the 19th International Conference on Data Engineering*, Banglore, India, 5th – 8th March 2003, pp. 337-348.

Ashish, N. and Knoblock, C. (1997a). Semi-automatic wrapper generation for internet information sources. *In Proceedings of Cooperative Information Systems*, pp. 160-169.

Ashish N. and Knoblock, C. (1997b). Wrapper generation for semi-structured internet sources. *In Workshop on Management of Semistructured Data*, pp. 8-15.

Badard, T. and Richard, D. (2001). Using XML for exchange of updating information between geographic information system, *Computers, Environment and Urban System 25(2001).* pp. 17-31.

Bayrak, C., Kolukisaoglu, H., Chung, H. and Talburt, J. (2002). Information harnessing on the world wide web. *In Proceedings of the 6th Biennial World Conference on Integrated Design and Process Technology,* 1, Pasadena : CA.

Bunge, M. A. (1977). Treatise on Basic Philosophy: *Ontology I: The Furniture of the World*, Reidel, Boston.(3).

Bunge, M. A. (1979). Treatise on Basic Philosophy: *Ontology II: A World of System.* Reidel, Boston.(4).

Castel, F. (2002). Ontological Computing. *In Communications of the ACM,* 45, (2), pp. 29-30.

Dieng, R., Corby, O., Giboin, A. and Ribiere, M. (1999). Methods and Tools for Corporate Knowledge Management. *International Journal of Human-Computer Studies (IJHCS)*, 51, pp. 567–598.

Egyedi, T. M. and Loeffon, A. G. A. J. (2002). Succession in standardization: grafting XML into SGML, *Computer Standard & Interface*, 24, (4).

Embley, D., Campbell, D., Jiang, Y., Ng Y., Smith, R., Liddle, S. and Quass, W. (1998). A conceptual-modeling approach to extracting data from the web. *In Proceedings of the 17th International Conference on Conceptual Modeling,* pp. 78-91.

Embley, D. W., Campbell, D. M., Smith, R. D. and Liddle, S. W. (1998).Ontology-Based Extraction and Structuring of Information from Data-Rich Unstructured Documents. *In Proceedings of the 1998 ACM CIKM International Conference on Information and Knowledge Management*, Bethesda, Maryland, USA, 3-7 November 1998, pp. 52-59.

Fensel, D. (2001). Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Heidelberg : Germany.

Frederico, T., Fonseca. and Egenhofer, M. J. (1999). Ontology-Driven Geographic Information Systems. In 7th ACM Symposium on *Advances in Geographic Information Systems Kansas City.* MOC. Bauzer Medeiros (ed).

Gibbins, N., Harris, S., and Shadbolt, N. (2003). Agent-based semantic web services. *In The Twelfth International World Wide Web Conference (WWW2003)*, Budapest, Hungary, ACM Press.

G′omez, M., Abasolo, C., and Plaza, E. (2001). Domain-independent ontologies for cooperative information agents. *Lecture Notes in Artificial Intelligence*, 2128, pp. 118–129.

Gruber, T. R. (1993). Towards Principles for the Design of Ontologies Used for Knowledge Sharing: Knowledge System Laboratory. Stanford University.

Guarino, N., and Giaretta, P. (1995). Ontologies and Knowledge Bases: Towards a Terminological Clarification. *In: N. J. I. Mars (ed.), Towards Very Large Knowledge Bases*, IOS Press, pp. 25-32.

Haas, L. M., Kossmann, D., Wimmers, E. L., and Yang, J. (1997). Optimizing queries across diverse data sources. In Proceeding of the 1997 Intl. Conf. on *Very Large Data Bases*, pp. 276–285.

Harold, E. R. (1998). XML: Extensible Markup Language, India : *IDG Books World Wide Inc.*

Hewett, K.A.(2000).An Integrated Ontology Development Environment for Data Extraction. Master's thesis, Department of Computer Science, Brigham Young University, Provo, Utah.

Levy, A., Rajaraman, A. and Ordille, J. J. (1996). Querying heterogeneous information sources using source descriptions. *In Proceeding of the 1996 Intelligent Conference on Very Large Data Bases*, pp. 251–262.

Magnin, L., Snoussi, H., Pham, V. T., Dury, A. and J.-Y. Nie.(2002). Agents Need to Become Welcome. *In Proceedings of the 3rd International Symposium on Multi-Agent Systems, Large Complex Systems, and E-Businesses (MALCEB'2002).* Erfurt/Thuringia, Germany.

Mika, P., Iosif, V., Sure, Y. and Akkermans, H. (2004). Ontology-based Content Management in a Virtual Organization. *Handbook on Ontologies 2004*, pp. 455-476.

Mohammadian, M. (2001). Intelligent Data Mining and Information Retrieval Retrieved 15 March 2004 from World Wide Web for E-Business Applications. http://www.ssgrr.it/en/ssgrr2002w/papers/230.pdf.

Molina, H. G., Papakonstantinou, Y., Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J. D. and Widom, J. (1997). The TSIMMIS project: Integration of heterogenous information sources. Journal of *Intelligent Information Systems*, 8, (2), pp. 117–132.

Myllymaki, J. (2C01). Effective Web Data Extraction with Standard XML Technologies. International World Wide Web Conference, *In Proceedings of The Tenth International Conference on WWW*, ACM Press New York, USA, pp. 689-696.

Nunamaker, J. F., Chen, M. and Purdin, T. D. M. (1991). System Development in Information Systems Research, Journal of *Management Information Systems*, 7, (3), pp. 89-106.

On-To-Knowledge. IST project IST-1999-10132 On-To-Knowledge, 1999. Retrieved 10 Mei 2004 from http://www.ontoknowledge.org/

OntoWeb,2002: OntoWeb. IST project IST-2000-29243. Retrieved 19 April 2004 from http://www.ontoweb.org

Sabuget, A. and Azavant, F. (1999). Building light-weight wrappers for legacy: Web data-sources using WW4F. *In Proceedings of the International Conference on Very Large Databases (VLDB'99)*, pp. 738-741.

Snoussi, H., Magnin, L. and Nie, J.-Y. (2002). Toward an Ontology-based Web Data Extraction. The AI-2002 Workshop on Business Agents and the Semantic Web (BASeWEB), *AI 2002 Conference (AI-2002)*, Calgary, Alberta, Canada.

Tijerino, Y. A., Embley, D. W., Lonsdale, D. W. and Nagy, G. (2003). Ontology Generation from Tables. In 4th International Conference on *Web Information Systems Engineering (WISE 2003)*, Rome, Italy.

Ullman, J. D., (1997). Information integration using logical views. *In Proceeding of 1997 Intelligent Conference on Database Theory*, pp. 19–40.

Wand, Y. (1989). A proposal for a formal model of objects. In W. Kim and F.H. Lochovsky, editors, Object-Oriented Concepts, Databases, and Applications, *ACM Press*, New York. pp. 537–559.