UNIVERSITI PUTRA MALAYSIA

TRANSFORMATION OF EXTRACTED KNOWLEDGE IN MALAY UNSTRUCTURED DOCUMENTS INTO AN INTERROGATIVE STRUCTURED FORM

FATIMAH SIDI.

FSKTM 2007 10

TRANSFORMATION OF EXTRACTED KNOWLEDGE IN MALAY UNSTRUCTURED DOCUMENTS INTO AN INTERROGATIVE STRUCTURED FORM

By

FATIMAH BINTI SIDI

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

SEPTEMBER 2007

** I admonish you, lest you be one of the ignorant * (Qur'an 11:46)*

Knowledge is a light that leads to wisdom.

It is life for one's soul and fuel for one's character.

**And say: "My Lord! Increase me in knowledge * (Qur'an 20:114)*

If you desire happiness,

then seek out knowledge and enlightenment,

and you will find that anxiety, depression, and grief will leave you.

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

TRANSFORMATION OF EXTRACTED KNOWLEDGE IN MALAY UNSTRUCTURED DOCUMENTS INTO AN INTERROGATIVE STRUCTURED FORM

By

FATIMAH BINTI SIDI

September 2007

Chairman: **Associate Professor Hj. Mohd. Hasan Selamat**

Faculty: **Computer Science and Information Technology**

The availability of knowledge discovery operation helps to extract valuable information and knowledge in large volumes of data in structured databases. However, a large portion of the available information is not in structured form but rather collections of text documents in unstructured format, which also implies to Malay unstructured documents. Therefore, structuring characteristics must be imposed to unstructured documents in order to transform information available in unstructured documents into knowledge. A new approach has been established to transform extracted knowledge in Malay unstructured document by identifying, organizing, and structuring them into interrogative structured form. Its architecture is developed based on the implementation of (i) interrogative knowledge identification; (ii) interrogative contextual information; and (iii) interrogative knowledge organization and

structuring with Malay knowledge representation by concepts. It utilizes the Malay language corpus; interrogative theory; as well as object-oriented, ontology, and database model. The research involves system development based on architecture of the Malay*i*K-Ontology, which is being measured by quantitative retrieval performance using the recall and precision metrics. The development of the Retrieval Interrogative Ontology Analysis Application is used to verify fitness of task for the functionalities and usefulness on the utilization of interrogative contextual information with color coding supplement, additional information annotation, and Malay knowledge representation by concepts. A number of experiments are carried out to quantify the accuracy of knowledge extracted. The Malay*i*K-Ontology is tested by using stratified random sampling drawn from various sources of Malay unstructured documents such as news, e-mails, articles, magazines, and texts from children story books. The results of the experiments have proved that the approach of Malay*i*K-Ontology performed well as compared to knowledge extracted manually done by an expert. The results of questionnaires evaluation on the Retrieval Interrogative Ontology Analysis Application have shown good achievement in understanding the main point of the unstructured document easily and clearly. This is to improve better understanding the process of making sense of information into knowledge, maintaining the meaning of the information and gaining the interpretation of the identical knowledge in unstructured document which facilitate identical knowledge perceived by different people.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

TRANSFORMATION OF EXTRACTED KNOWLEDGE IN MALAY UNSTRUCTURED DOCUMENTS INTO AN INTERROGATIVE STRUCTURED FORM

Oleh

FATIMAH BINTI SIDI

September 2007

Pengerusi: **Profesor Madya Hj. Mohd. Hasan Selamat**

Fakulti: **Sains Komputer dan Teknologi Maklumat**

Dengan adanya operasi penemuan pengetahuan, ianya telah membantu perolehan informasi dan pengetahuan yang berharga dalam saiz data yang besar dalam pangkalan data yang berstruktur. Walau bagaimanapun, sebahagian besar daripada informasi yang ada adalah berbentuk tidak berstruktur tetapi lebih kepada kumpulan dokumen-dokumen teks yang bentuknya tidak berstruktur, begitu juga dengan dokumen Melayu. Oleh itu, ciri-ciri pengstrukturan perlu dilakukan ke atas dokumen yang tidak berstruktur bagi membolehkan informasi yang terdapat dalam dokumen yang tidak berstruktur tersebut ditukar kepada pengetahuan. Satu pendekatan baharu telah diwujudkan untuk menukar pengetahuan yang diperolehi dalam dokumen Melayu yang tidak berstruktur dengan mengenal pasti, mengorganisasi dan menstruktur pengetahuan yang diperolehi ke dalam

struktur berbentuk interogatif. Pembangunan seni binanya adalah berdasarkan kepada pelaksanaan pengenalpastian pengetahuan interogatif, informasi mengikut konteks interogatif, dan pengorganisasian dan pengstrukturan interogatif dengan perwakilan pengetahuan Melayu melalui konsep. Ia menggunakan korpus Bahasa Melayu, teori interogatif, serta model yang berorientasikan objek, ontologi, dan pangkalan data. Penyelidikan ini melibatkan pembangunan sistem berasaskan seni bina Malay*IK*-Ontology, yang penghasilannya diukur melalui prestasi dapatan semula kuantitatif dengan menggunakan metrik perolehan kembali dan metrik ketepatan. Pembangunan Aplikasi Analisis Dapatan Semula Interogatif Ontologi digunakan bagi menilai ketahanan tugasan. Ianya dinilai melalui fungsi dan manfaat berkaitan dengan penggunaan informasi mengikut konteks interogatif dengan penambahan pengekodan warna, tambahan anotasi informasi, dan dengan perwakilan pengetahuan Melayu melalui konsep. Beberapa uji kaji telah dijalankan untuk mengenal pasti jumlah ketepatan pengetahuan yang diperolehi. Malay*IK*-Ontology diuji dengan menggunakan pensampelan rawak strata yang sumbernya adalah daripada pelbagai dokumen Melayu yang tidak berstruktur seperti surat khabar, e-mel, artikel, majalah, dan teks daripada buku cerita kanak-kanak. Keputusan daripada uji kaji telah membuktikan bahawa pendekatan Malay*IK*-Ontology telah menunjukkan prestasi yang baik jika dibandingkan dengan pengetahuan yang diperolehi secara manual oleh pakar. Keputusan daripada penilaian soal selidik melalui Aplikasi Analisis Dapatan Semula Interogatif Ontologi telah menunjukkan pencapaian yang baik dalam

memahami perkara utama dokumen yang tidak berstruktur dengan mudah dan jelas. Ini adalah untuk menambah lagi pemahaman proses boleh terima informasi menjadi pengetahuan, pengekalan makna informasi dan mendapat tafsiran pengetahuan yang sama dalam dokumen yang tidak berstruktur bagi memudahkan pengetahuan yang sama diamati oleh orang yang berbeza.

ACKNOWLEDGEMENTS

In the name of *Allah*, the most merciful and most compassionate. Praise to *Allah s.w.t* for granting me strength, courage, patience and inspiration in completing this work. First and foremost, I owe deep gratitude to Assoc. Prof. Hj. Mohd. Hasan Selamat, Chairman of the supervisory committee for his constant guidance and continuous encouragement without which this work would not materialized. I am also thankful to co-supervisors, Assoc. Prof. Dr. Abdul Azim Abd. Ghani, Assoc. Prof. Dr. Hj. Md. Nasir Sulaiman for sharing their intellectual experience. I am also fortunate to have the support and guidance from Dr. Mokhtar Mohd. Yusof of Malaysian Ministry of Health for his support in the early chapters of my work. To Prof. Dr. Hj. Awang Sariyan from Academy of Malay Studies, Universiti Malaya, I am very grateful for his expertise and intellectual experience in verifying Malay language for this research. Last but not least, thanks to Assoc. Prof. Dr. Hamidah Ibrahim for her valuable advice.

I am also indebted to Universiti Putra Malaysia for the study leave, scholarship and allowances which enable me to pursue this research. My gratitude is also extended to my parents, family, colleagues and friends, thank you for sharing all the pains and gains throughout the years. My special thanks and gratitude to my beloved husband Abdul Rahman Hassan and my five children Ibrahiem Luqman, Aqilah, Mahirah, Aniqah and Aqidah for their understanding, caring, support and patience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CF	Concept Frames
CFG	Concept Frame Graph
DAML+OIL	Darpa Agent Markup Language
DML	Data Manipulation Language
FLogic	Frame Logic
HTML	Hypertext Markup Language
KDD	Knowledge Discovery in Databases
KIF	Knowledge Interchange Format
KRSs	Knowledge Representation Systems
NER	Name Entity Recognition
NLP	Natural Language Processing
OCML	Operational Conceptual Modeling Languages
OIL	Ontology Interface Layer
OKBC	Open Knowledge Base Connectivity
OOP	Object-Oriented Programming
OSM	Object-oriented System Model
RDF	Resources Description Framework
RDF(S)	RDF Schema
SHOE	Single HTML Ontology Extension
SQL	Structured Query Language
XML	eXtended Markup Language
XOL	XML-based Ontology Languages

CHAPTER 1

INTRODUCTION

1.1 Background

The difficulty of defining knowledge in unstructured documents is due to the paradox that knowledge resides in a person's mind and at the same time, it has to be captured, stored, and reported. For that, philosophers classify knowledge into knowing-that and knowing-how. Knowing-that is factual where data are stored in databases and facts can be recalled, processed, and disseminated. While knowing-how is actionable to do something, turning data into information and in turn into knowledge (Spiegler, 2003).

Recent advancements in computer technologies and databases have given many approaches to generate and extract knowledge. Among them is data mining and many people treat data mining as a synonym for another popularly used term, Knowledge Discovery in Databases (KDD). KDD is the field that is evolving to provide automated analysis solutions in extracting and generating knowledge. It is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns/models in data (Fayyad *et al.*, 1996). Moreover, it has emerged as a rapidly growing interdisciplinary field that merges together databases, statistics, machine learning and related areas in order to extract valuable information and knowledge in large

volumes of data. KDD has deeply transformed the methods to interrogate traditional databases, where data are in structured form, by automatically finding new and unknown patterns in huge quantity of data. Most previous work in knowledge discovery is concerned with structured, numerical, heterogeneous databases and data warehouses, that will be extracted from the operational (day-to-day processing) systems (Poe, 1996; Iiritano and Ruffolo, 2001; Kroeze *et al.*, 2003). Much work has been done in developing and building knowledge discovery systems by using KDD process where the source data are from operational database (Buchheit *et al.*, 2000; Chen *et al.*, 2001; Ho *et al.*, 2001; Tsai and Chen, 2001; Valafar and Valafar, 2002; Leung *et al.*, 2003).

However, structured data represent only a little part of the overall organization of knowledge; in fact, the major part of this knowledge is incorporated in textual documents. For example, available business data are captured in text files that are not structured, e.g. memoranda and journal articles that are available electronically (Fayyad *et al.*, 1996; Iiritano and Ruffolo, 2001; Kroeze *et al.*, 2003). A large portion of the available information does not appear in structured databases but rather in collections of text articles drawn from various sources (Feldman, 1999). Thus, the main concern here is to dig knowledge from the available vast amount of textual documents.

1.2 Problem Statement

It is estimated that 90% of electronically available material is unstructured and the amount of unstructured textual documents, accessible through the web, intranets, news groups, etc. is enormously increased every year (Iiritano and Ruffolo, 2001). Hence, huge amount of unstructured documents are available on the web and intranets. The amount of information available to us is constantly increasing and our ability to absorb and process this information remains constant. "We are being drowned in information while being starved of knowledge and distracted from wisdom", taken from Dr. Norman Myers cited in Feldman (1999). Knowledge exists and is found everywhere (ubiquitous) in unstructured documents, so extracting knowledge in unstructured documents is essential.

Unstructured documents cannot be queried in simple ways. Therefore, knowledge contained in unstructured documents can neither be used by automatic systems nor could be understood easily and clearly by humans. Hence, identifying knowledge from unstructured documents to be easily realized and understood by humans is one of the most exciting areas to be explored.

Most work on knowledge discovery is concerned with structured databases. It is clear that this paradigm requires handling huge amount of information that is available in unstructured documents. To apply traditional knowledge

discovery or query operation on unstructured documents, it is necessary to impose some structures that will be rich enough to allow knowledge discovery operations techniques such as data mining to play their roles.

A structured database is a collection of related pieces of information stored electronically with structural description of the type of facts held in it. The most common model for a structured database is a relational model. It is used to represent any relationship between any collections of data that can be represented. It composes of tuples or records, and attributes or fields which unstructured document lacks. Embley *et al.* (1998a; 1998b; 1999a; 1999b) established and developed an approach of extracting information from unstructured documents and reformulating the information as relations in a database. The purpose of their work is to impose structure by establishing relations over the information contents of an unstructured document. The approach is based on data extraction ontology, a conceptual model instance that describes the data interest, including relationships, lexical appearance and context keywords. Later, Embley (2004) has extended the use of information extraction ontologies as an approach that leads to semantic understanding based on a foundation of Medows's definitions for data, information, knowledge and meaning. It is being reported that their approach generates precision ratios near 98% in extracting data on unstructured documents that are data rich. However, the work done does not attempt to extract "deep-level understanding" and does not depend upon complete sentences. The approach used is more appropriate for web pages

that publish information (like classified ads) which rarely contains complete sentences. Moreover, their approach is in extracting data on information of unstructured document; i.e., transforming information into data but does not focus on transforming information into knowledge. Unfortunately, nowadays a large part of knowledge is stored in an unstructured document or textual format which is usually written in complete sentences.

The growth in the number of unstructured documents written in Malay language is enormously available on the web and intranets. This triggers the need to investigate the availability of knowledge in Malay unstructured documents. The investigation looks on transforming information into knowledge. Hence, there is a need to identify the information in unstructured documents that have knowledge. The identified knowledge need to be extracted and then to be transformed into structured form by imposing structuring characteristic over the contents of the unstructured document. This is to enable the purpose of querying by using database standard Data Manipulation Language (DML), and increase understanding by humans on the main point of the unstructured document. At present, there are researches done in information retrieval in Malay language (Abu Ata, 1994; Ahmad, 1995; Tan, 1998; Abdullah, 2006). Unfortunately, these researches are focused on its information retrieval effectiveness but not on knowledge extraction.