BORANG PENGESAHAN STATUS TESIS

JUDUL:

Intelligent Searching System For Software Tutorials Using Natural Language Processing Technique . SESI PENGAJIAN: 2004/2005

Saya, Ho Shi Nee mengaku membenarkan tesis (PSM/Sarjana/Doktor Falsafah) ini disimpan di Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dengan syarat-syarat kegunaan seperti berikut:

1. Tesis adalah hakmilik Kolej Universiti Teknikal Kebangsaan Malaysia.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. ** Sila tandakan (/)

_______	SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
_______	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)
___/___	TIDAK TERHAD	

(TANDATANGAN PENULIS)

Alamat tetap:
252, KG BARU, 08100 Bedong,
Kedah Darulaman.

Tarikh :
22 Nov 2005

(TANDATANGAN PENYELIA)

Nama Penyelia:
Mdm. Halizah Basiron

Tarikh :
22/11/05

INTELLIGENT SEARCHING SYSTEM FOR SOFTWARE TUTORIALS USING NATURAL LANGUAGE PROCESSING TECHNIQUES

HO SHI NEE

This report is submitted in partial fulfillment of the requirements for the Bachelor of Computer Science (Software Development)

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
KOLEJ UNIVERSITI TEKNIKAL KEBANGSAAN MALAYSIA
2005

DECLARATION

I hereby declare that this project report entitled

INTELLIGENT SEARCHING SYSTEM FOR SOFTWARE TUTORIALS USING NATURAL LANGUAGE PROCESSING TECHNIQUE

is written by me and is my own effort and that no par has been plagiarized without citations.

STUDENT : ShiNee Date : 23.11.05
(HO SHI NEE)

SUPERVISOR : Date: 23.11.05
(MDM HALIZAH BT BASIRON)

DEDICATION

To my beloved family members

ACKNOWLEDGEMENTS

First thank goes to Pn Halizah Basiron, as my final year project supervisor. Along these four months of development of PSM II, she has been dedicated to her role by providing her professional and invaluable advice, and guidance in the preparation and development of this project. Thank you for being tolerant and patient. This has been one of the main factors that have brought to the successful completion of this PSM II.

Also, I would like to stress a gratitude to KUTKM and all the committee members of the PSM. Final Year Project was a good platform for undergraduates to apply their knowledge, skills in programming and management.

Special thank is dedicated to my faithful course mates who happen to be my housemates as well, for their unreserved care and sharing. Their presence has brought much joy and their kindliness and helpful hands are greatly appreciated.

Beside them, I shall not forget to express my appreciation towards Miss Zeratul Izzah binti Mohd Yusoh and Mr. Wilson Wong who though did not directly involve in my project but did provide some guidance, and also Dr. Nasina Jigeesh, my former supervisor for providing the idea of constructing such project.

Last but not least, I would like to dedicate these humble pieces of work to my family, who have always been faithful and have also been the constant source of strength from the start till the completion of this period.

ABSTRACT

This final year project entitled with **Intelligent Searching System For Software Tutorials Using Natural Language Processing Techniques (ISS)** is developed from scratch by an undergraduate starting from the preliminary process to implementation phase. The ISS is designed to search software tutorials stored in local machine, disk or Internet. It adopts the natural language processing technique to parse user query. As most searching engines employ the keyword matching techniques, the ISS build a simple parser based on bottom-up approach to tokenize, categorize and find the constituents of a sentence. The understanding of the structure of sentence enables the program to extract the keyword from users. As the program does not build its own search engine, the program parses the snippet (a brief description of a document) from Google Search Engine and highlights the relevant documents. The program uses some simple grammar rules that are suitable for the use of the program.

ABSTRAK

Projek Sarjana Muda II yang dinamakan *Intelligent Searching System For Software Tutorials Using Natural Language Processing Techniques (ISS)* adalah satu projek yang dibangunkan oleh pelajar universiti, bermula dari fasa analisa hingga ke fasa implementasi. Projek ISS ini direkakan untuk mencari latihan perisian dalam komputer, disk atau perangkaian Internet. Program ini menggunakan teknik pemprosesan bahasa semulajadi untuk menganalisakan ayat pertanyaan pengguna. Kebanyakan enjin pencarian menggunakan perbandingan kata kunci dalam ayat pengguna. Manakala program ISS akan menggunakan pendekatan *bottom up parsing* untuk membahagi, mengkategori and mendapatkan jujukannya untuk membina satu ayat betul. Disebabkan program ini tidak membina enjin perncarian sendiri, ia akan mem*parse*kan *snippet* (satu deskripsi ringkas tentang dokumen) yang dipulangkan dari *Google Search Engine* untuk mencari *relevancy* and menwarnakan perkataan yang mendekati makna tersebut. Program ini juga akan mengubahsuaikan tatabahasa semasa untuk kegunaan program ini.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ATTACHMENTS

LIST OF ABBREVIATIONS

NO	ABBREVIATION	NAME
1.	CSC	Computer Software Component
2.	CSCI	Computer Software Configuration Item
3.	IT	Information Technology
4.	ISS	Intelligent Searching System For Software Tutorials Using Natural Language Processing Techniques
5.	KUTKM	Kolej Universiti Teknikal Kebangsaan Malaysia
6.	NLP	Natural Language Processing
7.	OO	Object-Oriented
8.	OOA	Object-Oriented Analysis
9.	OOAD	Object-Oriented Analysis and Design
10.	PSM	Projek Sarjana Muda (Final Year Project)
11.	GDS	Google Desktop Search
12.	S	Sentence
13.	D	Determiner
14	N	Noun
15	V	Verb
16	NP	Noun Phrase
17	VP	Verb Phrase
18	ISP	Internet Service Provider
19	VSS	Visual Source Safe
20	WWW	World Wide Web
21	NL	Natural Language

CHAPTER I

INTRODUCTION

1.1 Project Background

The proposed system to be developed is a standalone system entitled with "Intelligent Searching System for Software Tutorials using Natural Language Processing techniques", in short, ISS. The following sections of this chapter discussed the background of the natural language processing, the application to be developed, objective and scope of the system, the significance of the proposed project and lastly the conclusion of the chapter.

1.1.1 Information Seeking

The computer world grows rapidly and all growth of each field, regardless of technology, programming, operating system, database system, occurs almost relatively simultaneously daily. The IT knowledge workers, lecturers and students have to continuously pursue the knowledge. They purchase or borrow the relevant books from library, update themselves by reading IT magazine regularly, or surf from the world wide web, the global world where the information are plenty to be found. Usually, people tend to surf Internet for free tutorials for some knowledge.

Most information seekers search their using various search engines on the web, for instance Google, Yahoo Search Engine, AskJeeves, UpdatedSearch, MSN Search, Lycos and the list continues. As most engines search almost all types of contents, (marketing, entertainment, financial, tourism, politic, living styles, fashion, books, astrology, any research field and many others), the result displayed to users are large, even though the specific has been mentioned in the query.

These engines offer great facility for users as it covers all fields. Thus, some users may just browse the Internet without any purposes, just like flipping through a magazine and "channel-surfing" on television. According to Marchionini (1995), this type behaviour is known as Undirected Browsing. While another two types of behaviours are Semidirected Browsing and Directed Browsing, which occurs in more systematic and focused way.

The proposed system would target on users of having these two types of behaviours who intends to search some software tutorials. As the scope is limited to a small scale and natural language processing techniques applied, the program is to provide better result for those tutorials seekers.

1.1.2 Natural Language Processing

Natural language is the human language that is used to exchange information, make a statement or request. In a partially observable world, communication can help agents be successful because they can learn information that is observed or inferred by others. Thus the processing of natural language by a machine enables the users to deliver their message to the machine for desired result.

Since the last five decades, the natural language processing program has been one of the most important subfields of developers and researchers. In short, NLP is the program deals with natural language in some way or another.

As quoted from the website, *http://www.pcai.com/web/glossary/pcai_n_o_glossary.html#Natural_Language_Processing*, " The goal of Natural Language Processing (NLP) is to design and build a computer system that will analyze, understand, and generate natural human-languages." and " One of the easiest tasks for a NLP system is to parse a sentence to determine its syntax. A more difficult task is determining the semantic meaning of a sentence. One of the most difficult tasks is the analysis of the context to determine the true meaning and comparing that with other text."

However the journey towards constructing a system that understands human language is very difficult due to some problems, as listed in the following:

- ***Word Sense Ambiguity*** : many words has more than one meaning; the program should be able to select the meaning that makes the most sense in context
- ***Syntactic ambiguity***: the grammar for NL is not unambiguous
- ***Imperfect or irregular input:*** Foreign or regional accent and vocal impediments in contextual information

However the journey never reaches its end yet and the achievement is remarkable and encouraging at current stage. The attempts are still continued by many researchers.

As for this project, the attempt is carried out on one of the famous tool of Internet Surfers or information seekers on local machine.

1.2 Problem Statements

Data and information are stored for reference and use. The abundance of data and information does not guarantee the access of the data itself when the user fails to track the data down, especially when the database grows larger and larger each day.

Beside the exact match of the keywords, the search engines return large number of result without understanding the query the user wants.

Thus, users need efficient search engine that understand their requirement.

1.3 Objectives

The proposed project consists of several objectives, as the following:

- To develop a tutorials retrieval program (search program) that processes human request in natural language form
- To utilize the services provided by Google Search Engine for effective searching.
- To develop a simple grammar and parser for the use of the ISS program.
- To filter information retrieved for users using NLP techniques

1.4 Scope

This application to be developed is a stand-alone system that can be run on user's own computer. Connection to Internet is not a must but the accessibility to World Wide Web enables the accessing of the global documents.

The natural language that is to process user's request is English. The users may not necessary to enter a full sentence but a more descriptive way helps in reaching the required documents.

The target user is students, lecturers and software tutorial seekers, particularly those are keen in self learning.

The program integrates with external search engine (Google) to search for certain documents like Microsoft office documents, acrobat documents, web page and others which are index-able by the search engine.

The Intelligent Searching System processes users query for software tutorials in natural language form and retrieve/capture the relevant resources from hard disc (where the tutorials is stored), CD ROM and with the connectivity to Internet available, from the web pages.

1.5 Project Significance

The Intelligent Searching System is specially built for tutorials seekers, particularly Information Technology Students. It enables the users to access easily the required software tutorials from a huge database (World Wide Web, hard disc and CD Rom).

Having a limited scope, the program is more efficient to extract the meaning of the query and hence locate documents with higher relevancy. This could reduce the browsing time and faster accessing to the required documents.

1.6 Expected Output

A list of software tutorials is displayed with path to its location according to the users' specification in natural language.