R 5.7053

Information Retrieval from the World Wide Web

Fauziah Baharom

29 September 1995.

**This dissertation is a part requirement for
the MSc in Software System Technology,
Department of Computer Science,
University of Sheffield.**

UNIT KOLEKSI KHAS

DECLARATION

(To he completed by the student and inserted into the dissertation immediately after the contents page)

All sentences or passages quoted in this thesis from other people's work have been specifically acknowledged by clear cross referencing to author, work and page(s). I understand that failure to do this amounts to plagiarism and will be considered grounds to failure in this thesis and the degree examination as a whole.

Name (please use block capitals) FAUZIAH BAHAROM

Signed .... Date 29/09/95

Abstract

This dissertation is being presented as a partial requirement for the degree of Master of Software System Technology in the University of Sheffield. The thesis on title "Information Retrieval From The World Wide Web" is undertaken from May to September 1995.

The aim of this project is to design a system which uses the Web searching technology. When this system is completed , it can be implemented to do indexing and searching any interested document that stored in the Web. This application may give some benefits to its users especially to the Department of Computer Science, University of Sheffield.

Ideally, this system is developed by using one of the software engineering approaches called "Incremental delivery strategy". The project began with the feasibility study of the techniques used in Information Retrieval and also the components involved in the World Wide Web and continued with the process of requirement analysis. The designer chose to use the data flow diagram method in order to show the design of this system. After the design was done, the system was implemented by using the C programming language and aided by the World Wide Web Library (or known as *libwww*).

This dissertation describes the development stages of this system. Problems faced and suggestions to recover were presented as discussions. And other relevant information was attached as appendixes.

Dedication

Dedicated to my parents,

Hajjah Embon binti Mat Isa and
Allahyarham Haji Baharom bin Nik Soh.

Acknowledgements

I would like to express thanks to my supervisor, Dr. Robert J. Gaizauskas, for his consistent support and guidance throughout my dissertation work. Despite his busy schedules, Dr. Robert J. Gaizauskas has given me his time for comments and suggestions on this dissertation.

A lots of thanks to Mr. Hamish Cunningham for his help especially in technical problems.

Special thanks to the Universiti Utara Malaysia for their financial support. My sincere thanks also goes to the Department of Computer Science, University of Sheffield staffs and all my friends, who always given me ideas and supports to finish this dissertation.

To my sisters and brothers, without whose help and unending generosity none of this would have been conceivable, thank you.

Contents

CHAPTER I : INTRODUCTION.

CHAPTER 2 : BACKGROUND.

CHAPTER 3 : SOFTWARE ENGINEERING.

CHAPTER 4 : SYSTEM REQUIREMENT.

CHAPTER 5 : SYSTEM DESIGN.

CHAPTER 6 : IMPLEMENTATION AND TESTING.

CHAPTER 7 : DISCUSSION.

CHAPTER 8 : CONCLUSION.

REFERENCES.

APPENDICES.

Chapter 1. Introduction.

1.0 Introduction.

The aim of this project is to understand the Web search technology and use it to design and develop an information retrieval tool.

1.1 Description Of The Problem.

Today there are many types-of media that can be used to provide information through the world. And with the help of expertise in science and technology, now people-can easily get information from network services. This network's information service is known as World Wide Web. This is a rival network of the Internet which has similar functions. Users can switch from one to the other smoothly without noticing the difference.

As we know, the World Wide Web is a huge information system on the Internet that uses a standard protocol (HTTP) and a standard for describing the structure of the documents (HTML). In general, the information retrieval system from the World Wide Web can be divided into two main activities which are indexing and searching.

This dissertation is aimed to design a system to retrieve information from the World Wide Web by referring to the concept and the design of the existing information retrieval system like Mosaic, Netscape and Gopher. The system will be developed by using a high level language. At the, moment, this dissertation tasks will cover on the development of indexing process. In general, the dissertation work carried out involved :

- Understanding the components of World Wide Web (also called WWW or the Web) and Wide Area Information Servers (WAIS) and how they are work.

- To generate a client which can serve an indexing request. Basically the indexing request is provided by a user. Before executing the request, the client must identify the type of that indexing request either only for accessing a document from the local file system or loading a document from a remote site.

- To find the linkage of all documents. When the first document is accessed successfully, the client then will try to look for embedded links in it. If the document do not containing any parse, then the process of loading a document will be stopped otherwise it will be continued until all the linkage documents was loaded by the client.

- To send the loaded document to the indexing engine.

- To create an indexing engine. This program will be used to generate an index files and doing others task which related with this process like to generate the synonyms table and to find stop word in the indexing request.

The contents of the thesis is for internal user only

References

[Rijs79] Van Rijsbergen, C.J., *Information Retrieval* , Butterworth & Co (Publisher) Ltd., 1979.

[Beoh81] Beohm, B., *Software Engineering Economics* , Prentice-Hall, 1981.

[Salt83] Salton, G. and McGill, M.J., *Introduction to Modern Information Retrieval* , McGraw-Hill, Inc., 1983.

[Pres87] Pressman, R. *S., Software Engineering A Practitioner's Approach* , McGraw Hill, Inc., 1987.

[Salt89] Salton, *G., Automatic Test Processing : the transformation, analysis, and retrieval of information by computer* , Addison-Wesley, Inc., 1989.

[Your89] Yourdon, E., *Modern Structured Analysis* , Prentice-Hall International, 1989.

[Liu94] Liu, C., Peek, J., and et. *al, Managing Internet Information Services* , O'Reilly & Associates, Inc., 1994.

[Eage94] Eager, B., *Using the World Wide Web* , Que Corporation, 1994.

[Hand95] Handley, M., Crowcroft, J., *The World Wide Web* , UCL Press, 1995.

[Turl95] Turlington, S. R., *Walking The World Wide Web* , Ventana Press, 1995.

[Gaiz94] Gaizauskas, R., *Notes Formal Method* , 1994-l 995.

[Cowl94] Cowling, T., *Notes Software Project Management* , 1994- 1995.

[Laff94] Lafferty, H., *Notes Structured System Analysis and Design System* , 1994-1995.

Hypertext References.

[Hypr1] http://www.tamu.edu/global-info/searcheng-intro.html

[Hypr2] http://www.w3.org/hypertext/WWW/Library/User/Architecture/DesignModel.html

[Hypr3] http://www.w3.org/hypertext/WWW/Library/User/Architecture/ControlFlow.html

[Hypr4] http://www.w3.org/hypertext/WWW/Library/User/Architecture/Threads. html

[Hypr5] http://www.w3.org/hypertext/WWW/Library/User/Architecture/DataStructure.html

[Hypr6] http://web.nexor.co.uk/mak/doc/robots/