

CERTIFICATION OF APPROVAL

Integrated Filtered Web-Search Engine

By

Wan Nordiana Wan Abd Kadir

Dissertation Submitted to the Information System Programme

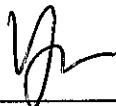
Universiti Teknologi PETRONAS

In partial fulfillment of the requirement for the

Bachelor of Technology (Hons)

(Business Information System)

Approved by,



(Ms Vivian Yong Suet Peng)

UNIVERSITI TEKNOLOGI PETRONAS

BANDAR SERI ISKANDAR, TRONOH,

PERAK DARUL RIDZUAN

JULY 2006

b

TK

5105.875

.157

W244

2006

i

Internet searching
in progress

CERTIFICATION OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this project, that the original work is my own except as specified in the references and acknowledgements, and that the original work contained herein has not been undertaken or done by unspecified sources or persons.



WAN NORDIANA WAN ABD KADIR

ABSTRACT

WWW has become one of the most important sources of information. WWW is not an indexed information warehouse where people easily look for specified data; it is instead a large collection of network of computers that contains the information. Finding information in the WWW can be as easy as it can be hard. Search engine was developed to assist users in searching information on the net. There exist a number of available effective search engine in the market nowadays but where human are concerns they always have something that they are not satisfied with. Mass information supplied to the users might get them exhausted as they browse through each and every one of the results returned. Even so, there were users who have the habits of only look at the top 10 of the results page and will go to another search engine if there still not satisfied with the information. This project aims to reduce users dilemma on mass information supplied as well as to combine the major search engines normally used by most users nowadays. The benefits are that users can have more results from various search engines with one single click without any redundant results.

ACKNOWLEDGEMENT

First and foremost I would like to thank my family especially my mother who had consistently given love, support, and encouragement in the course of completing this project.

I also would like to express my deepest gratitude and special thanks to my supervisor, Ms Vivian Yong Suet Peng who had monitored the progress of this project throughout the duration of the project. Her guidance means so much to me and her enthusiasm to make this project working, triggers my energy and enthusiasm as well that I've never known exist before. Under her supervision, I have experience lots of precious lessons that could be guiding me for the future undertakings.

The appreciation also goes to the other lecturer from AI group, who gave invaluable ideas and comments towards the better production of this project. Their critics, comments, honest opinion and suggestions have given me a better and clear picture of what it should take to make a high quality system.

Last but not least, I would also like to thank my friends who generously and cooperatively give comments and opinions towards this project.

TABLE OF CONTENTS

Chapter 1: INTRODUCTION	1-5
1.1. Background of Study	1
1.2. Problem Statement	2-3
1.2.1 Problem Identification	2
1.2.2 Significant of this Project	3
1.3. Objective and Scope of Study.	4-5
1.3.1 Main objective of this Research/Project	4
1.3.2 Scope of Study	4
Chapter 2: LITERARURE REVIEW & THEORY	6-13
2.1 Introduction	6
2.2 Literature Review	6-12
2.2.1 Web Linking and Web Integrating	6
2.2.2 Information Retrieval and Filtering	7
2.2.3Categorizing and Ranking the Result	10
2.2.4 Meta Search Engine Technology	11
2.3 Conclusion	13
CHAPTER 3:METHODOLOGY	14-30
3.1 Procedure Identification	14-29
3.1.1 System Development Process Model	14
Stage 1: Problem Identification.	15
Stage 2: Problem Analysis	16
Stage 3: System Design	17
Stage 4: System Development	19
Stage 5: System Testing	28
3.2 Tools Used.	29

CHAPTER 4: RESULT AND DISCUSSION	31-37
4.1 Results	31
4.2 Discussion.	35
4.2.1 Precision	36
4.2.2 IFWSE vs. MetaCrawler	38
CHAPTER 5: CONCLUSION & RECOMMENDATION.	40-41
5.1 Conclusion	40
5.2 Recommendation	41
REFERENCES	42
APPENDICES.	46

LIST OF FIGURES

- | | |
|------------|---|
| Figure 1.0 | Top search engine rate |
| Figure 3.0 | System Development Process Model |
| Figure 3.1 | System Architecture |
| Figure 3.2 | System Process Flow Diagram |
| Figure 3.3 | Pseudocode |
| Figure 3.4 | Connecting to remote search engine process flow |
| Figure 3.5 | Retrieving results process flow |
| Figure 3.6 | IFWSE Filtering process flow |
| Figure 3.3 | IFWSE Filtering Algorithm |

LIST OF TABLES

- | | |
|-----------|---|
| Table 3.0 | Hardware requirement |
| Table 4.0 | Comparison total results retrieved -1 |
| Table 4.1 | Comparison total results retrieved -2 |
| Table 4.2 | Precision table |
| Table 4.3 | Comparison of total result of IFWSE and MetaCrawler |

ABBREVIATIONS

IA	Intelligent Agent
LSI	Latent Semantic Indexing
LP	Natural Processing Language
FCM	Fuzzy Conceptual Matching
CFS	Conceptual Fuzzy Set
CLSI	Conceptual Latent Semantic Indexing
IR	Information Retrieval
FMQA	Fuzzy Modeling Query Assistant
FIS	Fuzzy Inference System
FOPC	first order predicate calculus
IFWSE	Integrated Filtered Web-Search Engine

CHAPTER 1

INTRODUCTION

1.1 Background of Study

Searching information on the web can be as extremely easy as it can be extremely difficult. This is because the WWW is not indexed like many library catalog or journal-article index. When we search on the web, we are not searching it directly but we are actually searching the web pages collected and indexed by a search tool from computers all over the world that contains the actual web pages. Still not entire web was covered by the search tools, but only portion collected by that index. Example of the search tools are Yahoo! Search, Google, AltaVista and etc.

The different types of search tools each have their own strengths and weaknesses. Depending on your information needs, one may work better for you than another. Search directories are hierarchical databases with references to websites. The websites that are included are hand picked by living human beings and classified according to the rules of that particular search service. Whereas, search engines use software to "crawl" the Internet in search of what you would like through the use of terms or keywords. Specialized databases are the hidden parts of the World Wide Web that are normally not found by regular search engines. [S. Chris-2005]

The study for this project namely, Integrated Filtered Web-Search Engine (IFWSE) is to enhance the searching strategies by utilizing the existing tools in the market nowadays. It is more on developing Meta-Search engine that have the abilities of filtering the results. The

search engine integrates the major search engines in the market which are Google, Yahoo! Search and MSN Search and is able to return the results from all those search engines to the integrated web search page without redundancy.

1.2 Problem Statement

1.2.1 Problem Identification

Back then no more than 10 years behind, it was said that a good search engine will have the ability to find any information on the web. Overtime, more websites developed and more information are available on the internet. And search engine is so popular by that time because of the simple processes to find information. These phenomena would force the search engine to handle millions of queries and information retrieving everyday. So it is very important for a search engine to have the capability of handling a large scale of queries and information retrieval [Liu-1998].

But today, with the massive information on the web and various kinds of websites offering knowledge to the surfers, it leads to such a tiresome and big burden to the users. They have to dig through all the search engine results in which by the end of the day turn out to be ‘junk result’. Normally the irrelevant results will wash out the results that the users are interested in. Yet with the advancement of the technology search engine nowadays normally give the best results and satisfied results to the users. The problem now here lay with the users habits themselves.

Some study shows that, users have the habit of looking at only the top 10 ranked search engine results [Liu-1998]. The iProspect Search Engine Branding Survey found that roughly 16 percent only look at a few entries of search results, and almost 32 percent read through the whole page. Only 23 percent of searchers go beyond to the second page, with the numbers dropping significantly for every page thereafter: first three pages (10.3 percent) and more

than three pages (8.7 percent). Almost 10 percent will read through the whole list of search results, unless it's dozens of pages. [G. Robyn- November 14, 2002]

If they still not satisfied with the results given they would normally go to the other search engine for some searching for the same keywords rather than trying new keywords in that same search engine. Other than that, even they are satisfied with what they searched for, users were always curious with what are the results from the other search engines so they tend to use more than two search engines that would results with more than two browsers to look through.

1.2.2 Significant/Benefits of This Project

i. Saving time in searching information in the internet

This search engine will simultaneously send queries to those three leading search engines and will return all the relevant results that have been filter up. User does not have to open more than one browser for different searches in other search engines.

ii. Retrieve up to top 50 results from each of the search engines

This filtered search engine retrieve up to top 50 results from each of the search engine and virtually there will be about 150 results altogether before filtered. Basically all the results presented to the users will be the most relevant.

iii. No duplicity of URL address

Information and data filter also will be done during the results retrieval.

iv. Easily used by anyone without a need to install

This search engine is design to be web-based engine where users do not need to install in order to use it. It is not like some of the search agent like Copernic where users have to install before they can use.

1.3 Objective and Scope of Study

1.3.1 Main objectives of this research/project:

i. Integrated search engines which can filter up the search results for any redundant URL

It is to integrate the leading search engine in the market which is Google, Yahoo! Search and MSN Search and the main idea is to filter out the search results for any redundancy so that users does not have to read the results twice or thrice. This will

ii. Eliminate duplicity in the results returned

Results returned to the users are thoroughly filtered

iii. Simultaneously searching on several popular search engine and retrieve the top 10 results of each search engine.

1.3.2 Scope of Study

IFWSE main objective is to integrate three major search engines in the market which are Google, Yahoo! Search and MSN Search. Those three were being selected as the sources because they are the leading search engines in the market today. Refer Figure 1.0 [S. Danny-January 24, 2006]. The remaining two major procedures are information retrieval on the web and information filtering for non-redundancies results.

Basically it will involve finding methods for web information retrieval and web information filtering that used by the other integrated search engines as well as search agents in the market and enhance the methods so that it will more convenient to the users.

Figure 1.0 shows the major search engines in the market and the percentage of users use the service. Based on the pie-chart above we can say that most users prefer to use Google, second is Yahoo and MSN got the third place in the attracting users. Obviously they are preferred because of their effectiveness in presenting the relevant results to the users.

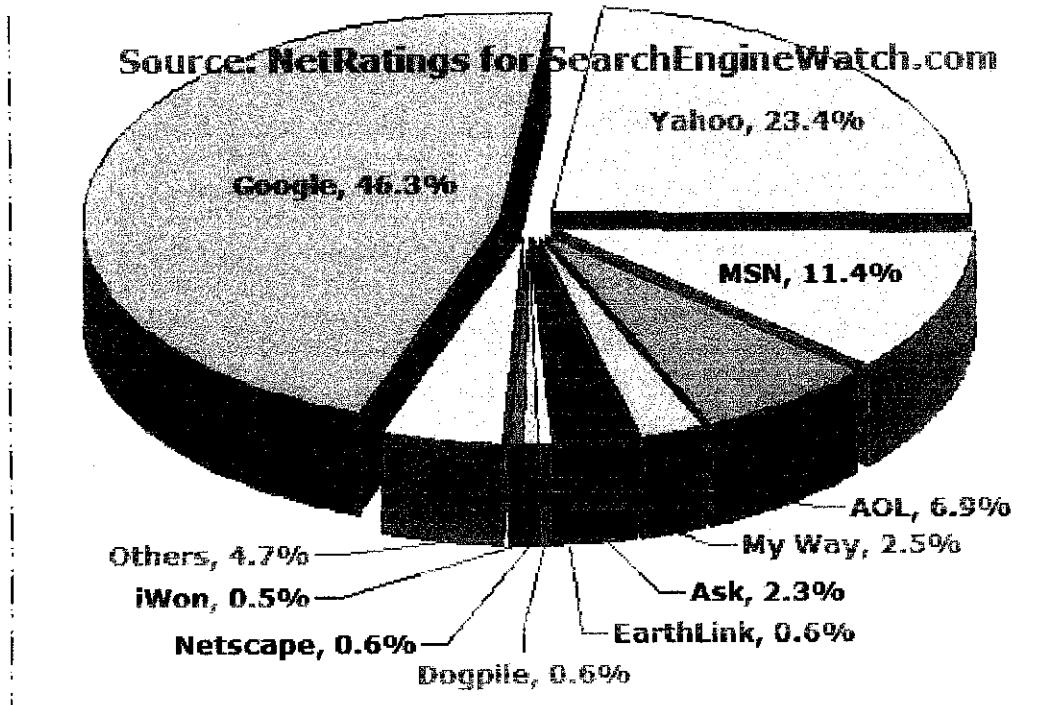


Figure 1.0: Top search engines rate [S. Danny- January 24, 2006]

CHAPTER 2

LITERATURE REVIEW & THEORY

2.1 Introduction

The simplest definition of an Intelligent Agent is ‘a software entity that assists people and acts on their behalf’ [Neal-1997]. According to Franklin and Graesser, autonomy is one of the most useful aspects in distinguishing IA from other types of software [Neal-1997]. Using IA in enhancing existing search engine would enable multi-task performing, intelligent search as well as autonomous entity which performing the entire task on the user’s behalf. The central task for the most of the search engine can be summarized as:

- “ 1) Query or user information request – do what I mean not what I say!
2) Model for Internet, Web representation
3) Ranking or matching function – degree of relevance, recall, precision, similarity and etc” [Neal-1997].

2.2 Literature Review

2.2.1 Web Linking and Web Integrating

According to Daniel, a link is simply a connection between the content of two different files (or between different parts of a single file) [Regina-2004]. Links in the website might lead to a different page of the website or to a page that’s from the other web site which is from the other computer. There are two types of linking which are Hypertext and Image.

Linking the page is not as complicated as integrating the web (search engine). All the web page has different layout and different structure of coding. A method has been

developed where the linking of the web page of the search engine will be made from its URL address [**Regina-2004**]. After the linking process, the searched keyword will be submitted for searching process and after all the information has been gathered it will be stripped to the Integrated Search Engine result's page [**Regina-2004**]. All this processes will be guided by a pseudocode.

2.2.2 Information Retrieval (IR) and Information Filtering

According to Tg. Mohd in [**T.M.T. Sembok-2003**], IR is concerned with the determining and retrieving of information that is relevant to the information need as expressed by his request and translated into a query which conforms to a specific information retrieval system (IRS) used. [**T.M.T. Sembok-2003**]

G. Michael Youbblood emphasized that, there are gaps in human-computer interface which leads to conflicts between the way people query information and the way computers store information [**G. Michael Youngblood-1999**]. Based on these he suggests that future web searching should emphasize in the Natural Language Processing (NLP) which uses the conceptual or semantic basis that allow the agent to search for ideas and not just words [**G. Michael Youngblood-1999**]. This approach is also being emphasized by the author in [**NLP-2005**] where the NLP can be used by using Latent Semantic Indexing (LSI). This methods can capture terms associations in documents where it is more likely a human behavior which computer has none. The technique could improve the search engine to do the searching more intelligently.

Basically the queries are treated as independent keywords or unstructured collections of keywords or terms which are generally assume to be statistically independent. In order to achieve a more accurate representation of documents and queries, the simple keywords

representation should be replaced by a knowledge representation such as semantic networks, logic, frames or production system [**T.M.T. Sembok-2003**]. NLP using logic in the form of first order predicated calculus (FOPC) to represent the contents of documents and queries was proved to be an effective way to improve the better understanding of the search engine with the human query or language [**T.M.T. Sembok-2003**]. The nouns or phrases are translated into predicate calculus for the computer to translate.

Masoud Nickravesh proposed that using Conceptual Fuzzy Set (CFS) model is very useful in enhancing information and knowledge retrieval through conceptual matching of both text and image [**N. Masoud-2003**]. The CFS model a.k.a. Fuzzy Conceptual Matching based on Human Mental Model is an integrated framework of clarification dialog, user profile, and context and ontology techniques of information retrieval. In the CFS model, the techniques used are conceptual matching of text, terms similarity, and fuzzy ontology [**N. Masoud-2003**].

The terms similarity which is based on Conceptual Latent Semantic Indexing can be constructed from the collection of text documents. Using all those CLSI, personalization and user profiling can help in query refinement, providing suggestion and also ranking the information. The conceptual matching of text is used to have the query selected doesn't need to be in exact matching with the decision criteria which is more human-like-behavior. The Fuzzy Conceptual Matching (FCM) can be used for constructing fuzzy ontology or terms related to the context of the query and search to resolve the ambiguity and imprecision of the concept describes by both textual and image information [**N. Masoud-2003**].

Fuzzy Inference System (FIS) can be used to process queries as well as the user profile created by the agent [**S. Vrettos, A. Stafylopatis-2001**]. The authors of the journal used several modules for their agent architecture which are Indexing Module, Profiling Module, Interface Module, Information Retrieval Module and Information Filtering Module.

There is short-term profile, long-term profile as well as the integrated profile built for this purpose. The profiling set is built based on the page that has been rated by the users, by number of times the page visited and so on. Interface module is used to determine the mode of the user intention either the “working on a project “or the “no specific goal” mode. For information retrieval and information filtering module, the searching will be based on the mode selected by the users and the mode will determine either the short-term profile or long-term profile will be used in the FIS [**S. Vrettos, A. Stafylopatis-2001**]. So basically, this will need more input from the users and will improved overtime when the profile is everyday improved. Generally the search engine will work better if the user put more information or make the queries more specific on a certain subject. So, the methods used in [**S. Vrettos, A. Stafylopatis-2001**] will help a lot in making the queries more specific to the user’s need and interest.

According to the authors in [**R.I. John, G.J. Mooney-2001**], using the combination of user modeling and fuzzy logic also know as. Fuzzy Modeling Query Assistant (FMQA) which modifies a user’s query based on a fuzzy user model proved to be better on getting the relevant information. FMQA employed the knowledge about users to modify the queries before sent out to the search engines. Knowledge about users is gotten from the questionnaires answered by the users – which can be used as the model of the user’s experiences and knowledge of the WWW. This can solve the vagueness, ambiguity, irrelevancy and redundancy problems faced by the IR in general [**R.I. John, G.J. Mooney-2001**].

2.2.3 Categorizing and ranking the Result

Search engine will sort through the millions of pages it knows about and present you with ones that match your topic. The matches will even be ranked, so that the most relevant ones come first.

They follow a set of rules, known as an algorithm which is unique from each other amongst the search engine as well as some general rules. One of the main rules in a ranking algorithm involves the location and frequency of keywords on a web page. Search engines will also check to see if the search keywords appear near the top of a web page, such as in the headline or in the first few paragraphs of text. They assume that any page relevant to the topic will mention those words right from the beginning. Frequency is the other major factor in how search engines determine relevancy. A search engine will analyze how often keywords appear in relation to other words in a web page. Those with a higher frequency are often deemed more relevant than other web pages [S. Danny- July 31, 2003].

In the [N. Masoud-2003] issue of ranking the result also being discussed, Masoud Nikravesh proposed using the Conceptual Latent Semantic Indexing (CLSI), together with personalization as well as user profiling. The user profile is automatically constructed from text document collection and can be used for query refinement and provide suggestions and for ranking the information based on pre-existence user profile.

According to document [S. Fabrizio-2004] the highly effective technique in ranking page is by using PageRank Technique which is applied in the Google search engine [S. Fabrizio-2004]. The PageRank of a page is computed by weighting each hyperlink proportionally to the quality of the page containing the hyperlink. To determine the quality of a referring page, they use its PageRank recursively [S. Fabrizio-2004]. In [B. Sergey and P. Lawrence] PageRank was assumed as a model of user behavior. An intuitive justification is

made where a page can have a high PageRank if there are many pages that point to it, or if there are some pages that point to it and have a high PageRank [B. Sergey and P. Lawrence]. So in Google results page, the top ranking results will be the results that has the highest number of page that point to that page.

2.2.4 Meta Search Engine Technology

Meta-search engines do not own a database of Web pages; they send your search terms to the databases maintained by search engine companies. "Smarter" meta-searcher technology includes clustering and linguistic analysis that attempts to show you themes within results, and some fancy textual analysis and display that can help you dig deeply into a set of results. However, neither of these technologies is any better than the quality of the search engine databases they obtain results from [B.Joe – 2005]. There are quite number of meta-search in the market already such as Dogpile, Mamma, MetaCrawler, Kartoo and Pandia search engine. All these have different types of ranking algorithm and retrieving techniques and retrieve results from various search engines.

Example of recently published meta-search engine:

The MetaCrawler works by querying a number of existing, free search engines, organizes the results into a uniform format, and displays them. A Fast Search produces results the quickest. After a few seconds this search method will bring up a new page filled with links to information related to your keywords (called "hits"). Alternatively, the Comprehensive Search button may be used. This will result in a longer search that produces more hits.

The MetaCrawler operates in two general modes: Normal Mode and Verification Mode. In Normal Mode, the MetaCrawler reports results immediately after retrieval from the remote search engines. In Verification mode the MetaCrawler loads and verifies each

reference to ensure the validity of the data. Thus the data returned is of much higher quality. MetaCrawler ranking uses Service Vote Rankings method to rank its results. It combines the confidence scores given to each reference by the services that return it. Thus, when the MetaCrawler returns a reference, it sums the scores given by each service and presents them in a "voted" ordering." **[B.Joe – 2005]**

There are numbers of Intelligent Search Agent in the market which act similarly to Meta-Search where it can search simultaneously several search engines at one time. Example is Copernic 2001. Copernic features a search wizard, the ability to search using a question or keywords, keyword highlighting in results and Web pages, a detailed search history, automatic software updating and many useful search management functions **[Copernic-2005]**. Combining robustness and scalability, this technology retrieves and indexes data wherever it is found: on corporate intranets, company servers, and public Web sites. It makes use of advanced language and linguistic analysis technologies, resulting in unparalleled indexing precision **[Copernic-2005]**.

2.3 Conclusion

Basically the most used techniques in enhancing the information retrieval and filtering are personalization and user profiling. Determining the user profile and personalization for later used in the intelligent search engine methods seems very essential in retrieving the relevant information based on the user's need and interest as it provide the specification for the search engine to do the retrieval processes.. This is very obvious when every single previous research would include a profiling and personalizing the user in one or their methods in improving the information retrieval and information filtering. These two techniques help a lot in retrieving the most relevant information that the user might be interested.

Other than that, the fuzzy approach is very useful in handling the ambiguity and imprecision results, the common problems that generally faced by most search engines. The currently researched method is NLP which promotes the implementation of a human-like behavior to the search agents. Basically search engines technologies nowadays are advanced enough that it can think similarly like human do and more than that it can perform the task without human intervention as we can see the capabilities of Copernic.

CHAPTER 3

METHODOLOGY

3.1 Procedure Identification

In this section, the procedure involves in developing the system will be discussed. This project basically uses Development Model derived from Waterfall Model. All the activities involved in each stage will be discussed in details.

3.1.1 System Development Process Model

Figure 3.0 illustrates the System Development Model phase for Intelligent Web-Search Agent. This model derived from the Waterfall Model Process [Marshall et al, 1994]. There are six stages of development phase for this system and it follows the concept of Waterfall Model where it applies the linear sequential model process (each stage related to each other and happens one after another).

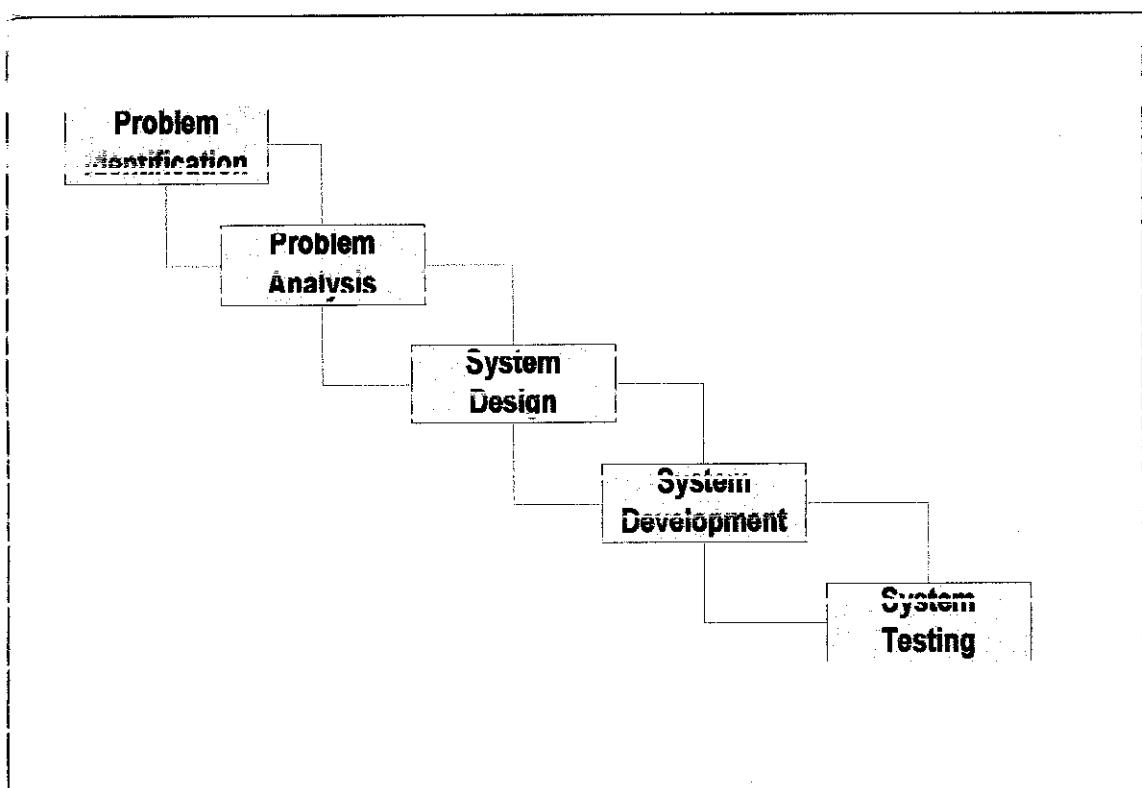


Figure 3.9: System Development Process Model

Stage 1: Problem Identification

Rapid development of websites and mass information available makes the internet a huge warehouse for information. Search engines became the popular search tools in information finding on the internet. More people prefer to use search engines as it was the most simple and easy to use as well as effective ways nowadays in information finding.

In this stage, some significant problems arouse due to the above situation were being acknowledged, identified and classified. When excessive information being supplied at once people would get tired in reading each of them, and also exhausted in determining the most relevant and useful information for them.

Other problem being identified was the search engine had to suffer due to the task of handling thousands of queries per minute and had to search through the unstructured nature of data on the. The problems also being identified lays at the search engine technologies that sometimes couldn't cope up with the excessive demands from million users. The unstructured nature of the information on the internet and the websites were not standardized in one format to help the search processes.

Nowadays there are numbers of intelligent search engine available like Copernic, but the drawback is they are still stand-alone system and need to be installed before using.

Stage 2: Problem Analysis

Basically the problem analysis was being done through literature review and articles studies as well as existing search engine studies. The main focus for the problem analysis is the relevancies of the results returned to the users. Most of the top search engines now like Google, Yahoo! Search and MSN Search are considered intelligent search engines. These search engines had applied the intelligent techniques in retrieving information like Latent Semantic Indexing, fuzzy approach, user profiling and etc.

Intelligent search engine like Copernic also being studied. This agent basically used many integrated features to enhance searching process like intelligent information retrieval, information filtering through queries refining, removing broken link and user profiling. This software was actually not a web based agent but the features were quite useful to eliminate any unnecessary link as well as save time in the searching process as it can be programmed and operate by itself. But the drawback of this intelligent search engine is it has to be installed before it can be used.

Other than intelligent searching software, writer also studied the nature of Meta-Search engine. Basically this project has the nature that was almost as similar as Meta-Search engine. This type of search engine has the capability to search various search engines at once and compiled the results in one integrated interface. There were quite lots of number of Meta-Search engine being developed such as Dogpile, Ask Jeeves, Meta Find, Meta Crawler and Mamma. All these have their features and performance that were quite similar with each other. Several of them filter out the results but there are some just display the whole page of various search engines in one pages. Most of the Meta search engine not included Google or MSN or Yahoo altogether but the Meta Crawler or Web Crawler do include them as the main search engine.

Stage 3: System Design

During this stage the system was designed conceptually using the process flow diagram to illustrate the function procedure of the system clearer. The system's input, output, external interactions, processes and procedure were all being identified during this stage.

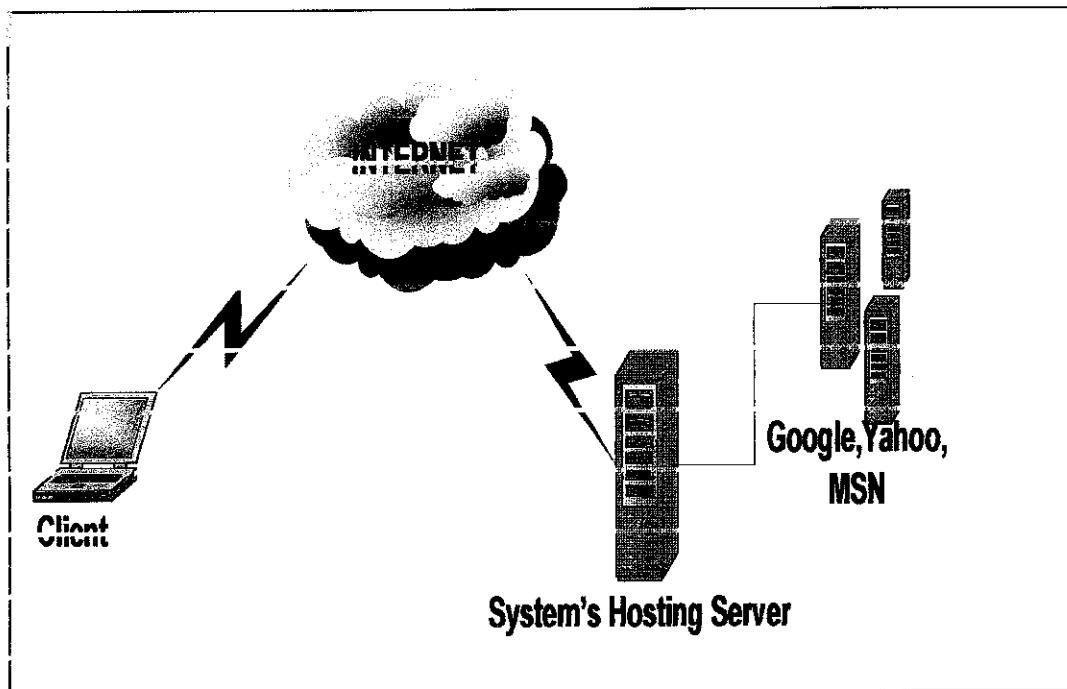


Figure 3.1: System architecture

Figure 3.1 illustrate the architecture of the system. In this diagram all the external entities that interact with the system being shown. This diagram also indicates the communication methods being used by the system as well as the external entities with each other.

Basically it shows that, this system will provide a real time information as it directly connected to remote search engine. End-users are connected with the system via Secure or Private Internet connection. The system was hosted on the hosting server that will be online 24-7 and the system's hosting server will be connected with the remote search engine via Internet connection.

Process Flow Diagram was to illustrate how the queries being manipulated to produce and present desired and relevant results to the user. The processes like sending queries to the external search engines, retrieving the results, filtering the results as well as ranking the results accordingly were being identified, defined and illustrated in Figure 3.2.

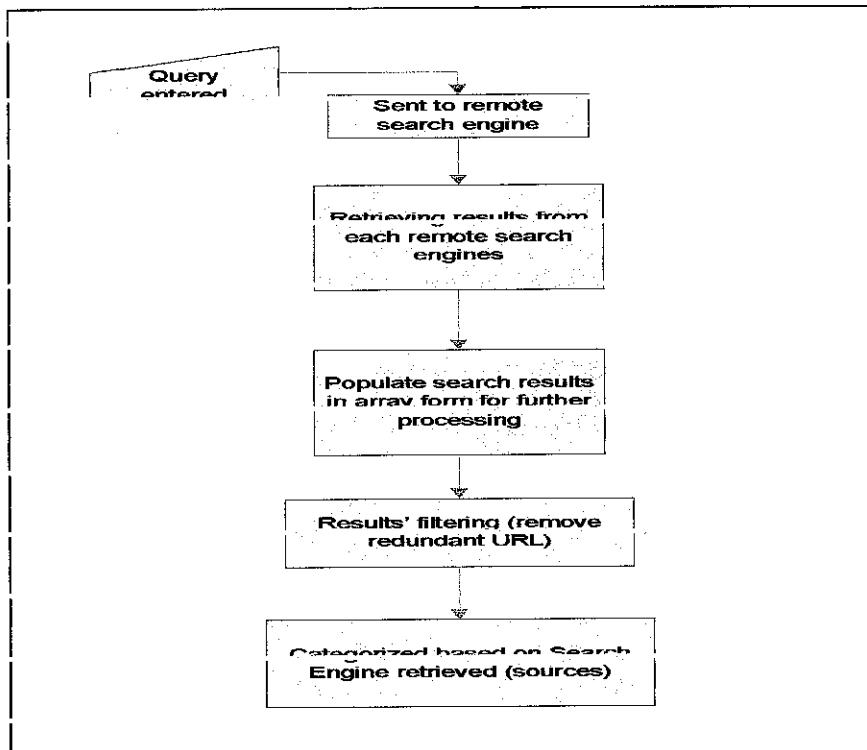


Figure 3.2: System Process Flow Diagram

Figure 3.2 illustrate the flow of the processes involved in this system. The first rectangular like shape indicate the input being entered from user. This input or queries will be submitted to the external search engines. Those search engines will search keyword submitted in the internet or web for any relevant and related topics. The accuracy of the retrieved results will depend on the external search engines effectiveness in retrieving information intelligently. That's why it is important to choose the best search engine for the platform search engines.

The results from the search engines will then be retrieved by the system. The IFWSE system will not retrieve all the results returned, only top 50 results from each of the search engines will be chosen. The results then will be filtered out for any irrelevant and duplicate link. The last process was to rank the results according to the user preferences or according to the most relevant to the user's interest. Only after the ranking process, the results will be displayed and presented to the user.

Based on the process flow above, the **pseudocode** then was written. The purpose of this pseudocode being written was to ease the task during the development stage as it clearly shows the function procedure of the system. Figure 3.3 below shows the pseudocode that was being derived from the process flow. Basically the pseudocode represents the processes of the system in details.

Submitting Query

Show the main interface

Set variable query as function values

Call the main function for sending queries to the remote search engines processes

Major Process that will perform results retrieval and parsing URL

Call the QueryFunction

QueryFunction return results retrieved from remote search engines

Converting results from strings to arrays

Filtering process will be on identical URL

If found same URL

 Mark current Google results also from get from Yahoo

 Delete current Yahoo results

 Re-order Yahoo results array

 Break;

If same URL not found

 Set pointer to next Yahoo result

 Go to next Yahoo result for comparisons

(Process will be between Google-Yahoo, Google-MSN, and Yahoo-MSN)

Populate the filtered results based on sources (retrieved from)

Output is displayed based on the categorized filtered results

Display the results

QueryFunction

Called from Main Process

Connecting to the remote search engine

Retrieve all top 50 results from each of the remote search engines

Stripped the header and footer of each of the results retrieved

Return the results to the Main Process

Figure 3.3: Pseudocode

Stage 4: System Development

The system started being developed during this phase. For this system, the writer divides the process into two which were system prototype development and the real system development continuing from the system prototype state.

The interface was developed using PHP scripting language, aided by Macromedia Dreamweaver MX. The Apache web-server was the local web-server. As for the prototype, MySQL was used as the database and phpMyAdmin was used to aid in interacting with the database. Interface designed was not a major concern for this system but still the interface is designed as simple as possible and users can easily adapt to the interface.

For the second phase, the real methods and techniques started being used to develop an intelligent system. Processes and procedures that were being focused were, connecting to the remote search engines, sending queries to the remote search engines, retrieving results from the remote search engines' results' page, filtering the results, ranking the results and finally displaying the results according to its ranks accordingly to the system's results' page. The mentioned lists of processes were basically the user defined functions contained in this system. Within the user defined functions there were number of php functions being used by the writer. Using PHP scripting language basically helps a lot in process of developing the system as there were lots of functions that were already written to the PHP developer. Along the way in explaining the process of developing each of the user defined functions, the existing php functions used will be mentioned as well.

1) Connecting to the remote search engine

Figure 3.4 below illustrate the process flow of connecting to remote search engine which are Google, Yahoo and MSN. This process was being performed when user click search button after entering query.

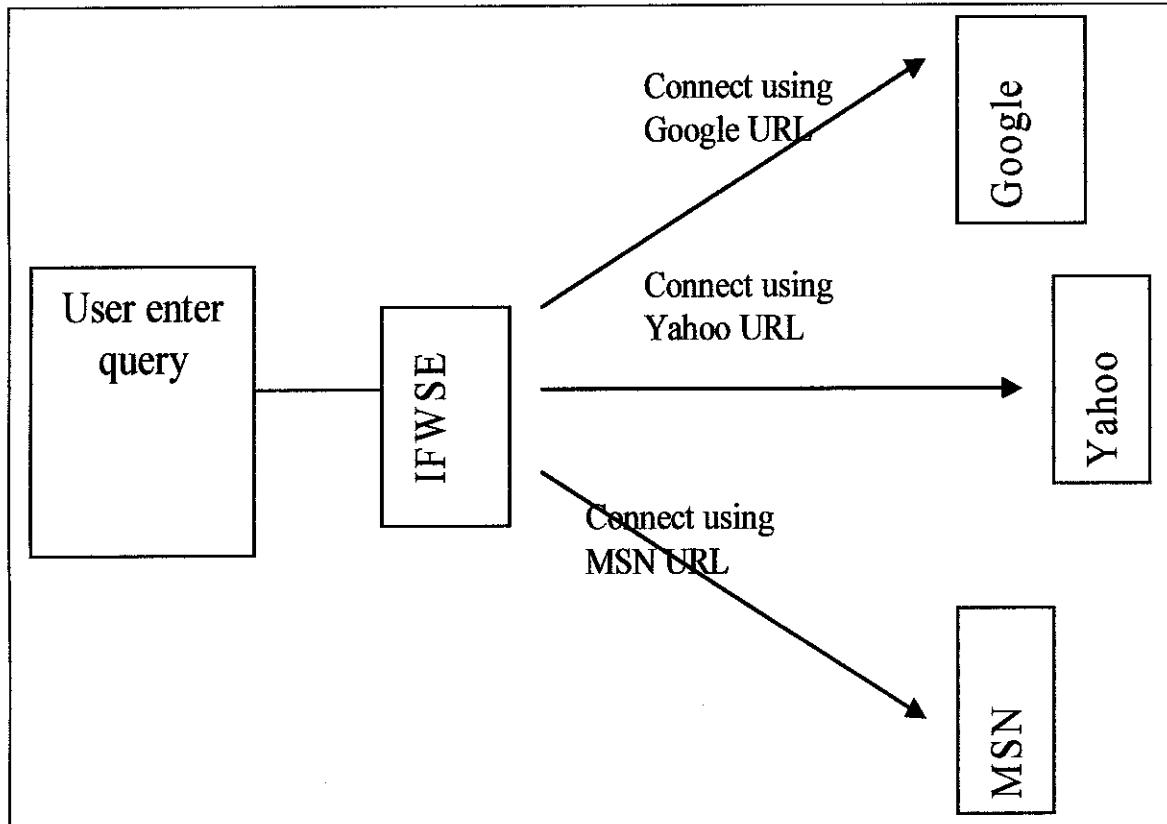


Figure 3.4: Connecting to remote search engine process flow

Query entered by users will be directed to the remote search engine. This IFWSE is connected to the search engine only when user click search button. This will trigger function that is responsible to do the connecting process. Each of the search engines is connected using separate function as they need to use different URL to connect to Google, Yahoo and MSN.

2) Retrieving results

Figure 3.5 below illustrate the process flow of retrieving results from remote search engine. This process was being performed after query entered by users submitted to the remote search engine and search engine responses by returning the search results.

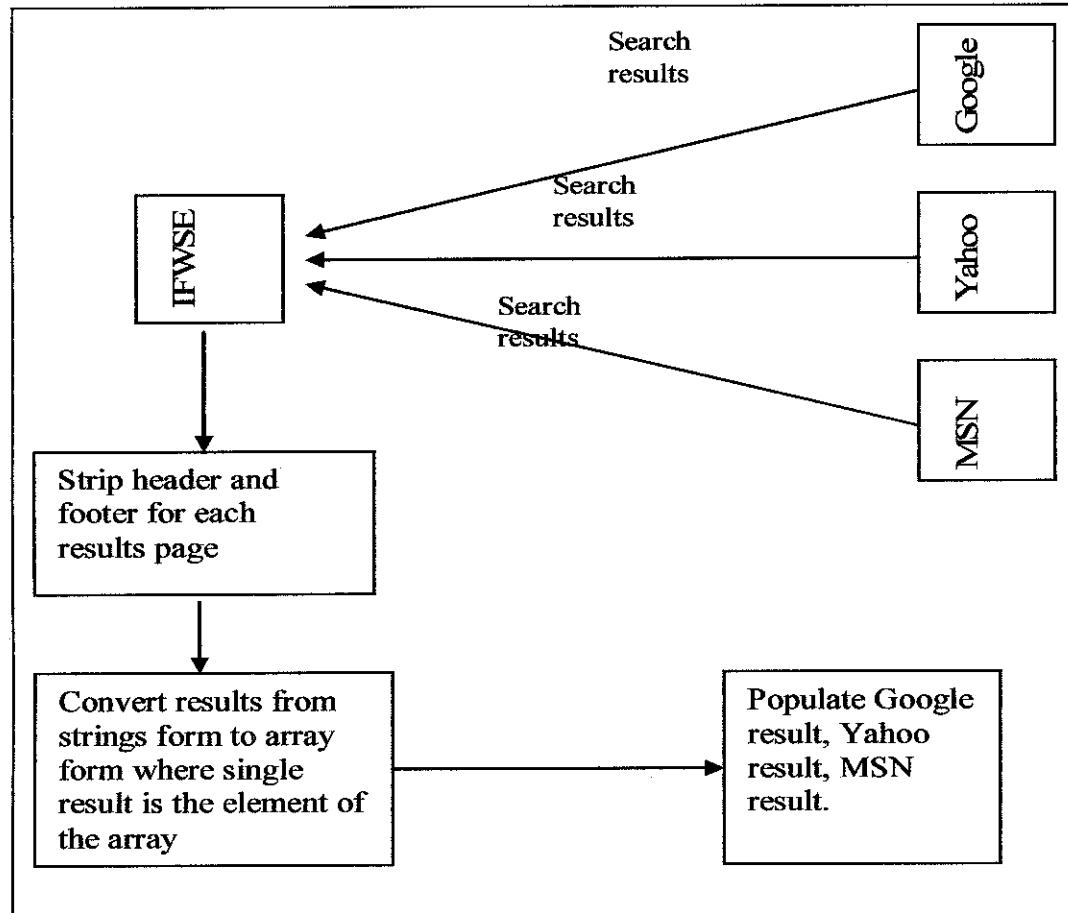


Figure 3.5 Retrieving results process flow

Function `file_get_contents` is used together with the supplied URL to retrieve the search results from each of the search engines. This function is built in function in php that enables information retrieval in remote file. The system basically retrieved the whole page where all the information in the form of string. Before the results can be any useful the header and footer need to be stripped away and only leave the results that would normally contain topic, description, URL and additional similar links. To get rid of the header and footer, string manipulation function is used, `explode` where it will divide the results into three separate parts which are header, body and footer.

Before filtering function being call, the results will be further processed from strings to array format. Using string manipulation function available in php like *explode* or *split* at certain line will separate each of the result into single result that contain all the usual information like Title, Description, URL and some other related links. To make the array much simpler and easier to handle the array is populated so that the elements are arranged in similar form for filtering function processes.

3) Filtering results processes for any redundancy

Figure 3.6 below illustrate the process flow of filtering results. Basically this process was being performed after all Google, Yahoo and MSN results being properly populated and converted into array.

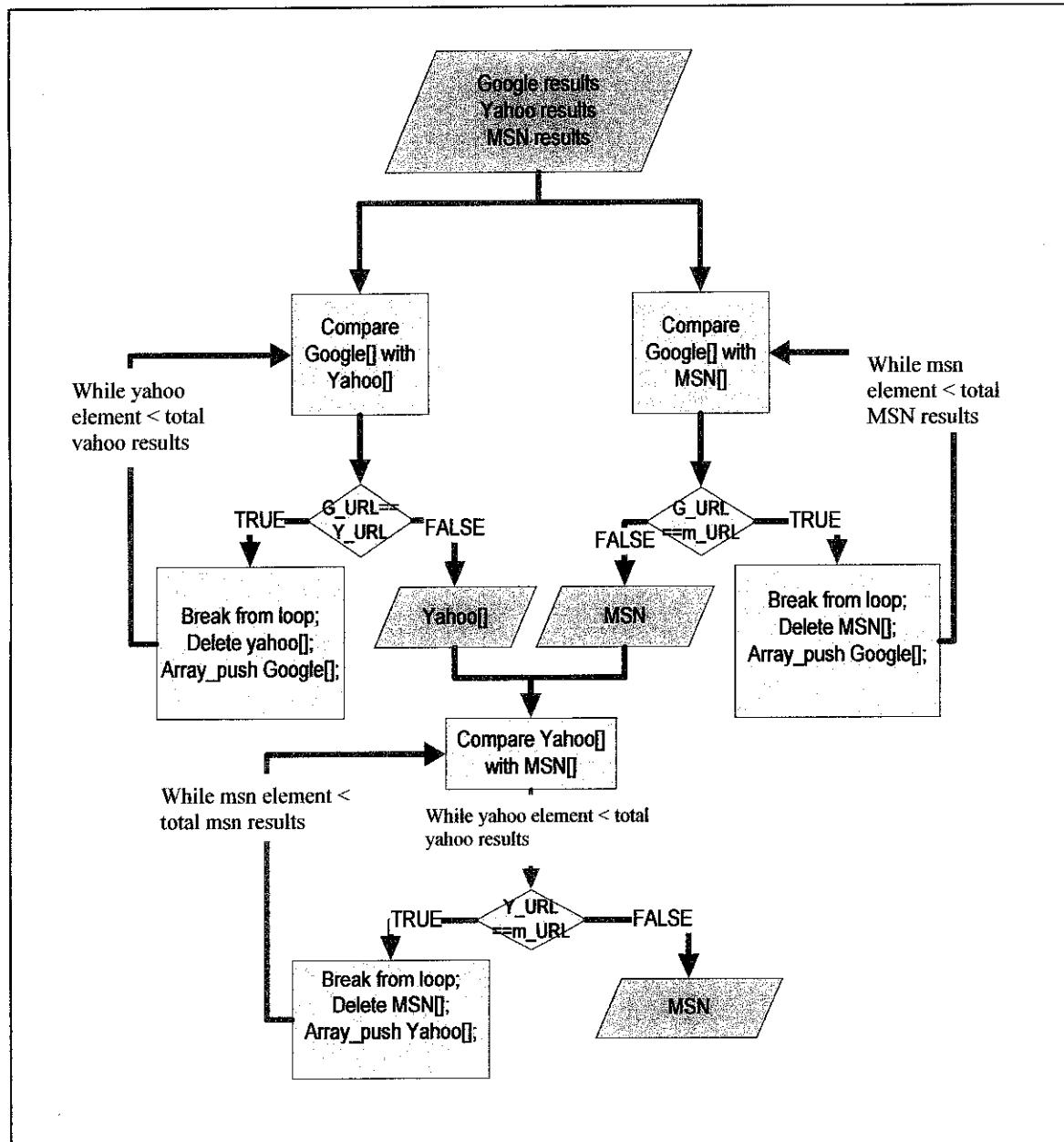


Figure 3.6: IFSWE filtering process diagram

Figure 3.4 are the filtering algorithm flow diagram and pseudocode. Both represent the same process but in different form of representation. The filtering process is done after the array of the results have been populated and arrange in the form where it is easier to do filtering and sorting processes.

The results from each of the search engine are being filtered step by step using the Google as the base search results and compared using URL to check any redundancy. For the start, Google results will be compared with the Yahoo results and comparison for Google and MSN will follow afterward. Then the remaining Yahoo and MSN results that have no similar URL with Google will be compared for any redundancy in URL.

The process will be recursive while looping till the maximum number of the results. If the similar URL found while comparing, the loop will be break and continue to the next result. One of the results that have similar URL will be deleted, like if Google result compared with Yahoo result and identical URL found, the Yahoo result will be deleted so that later on it will not be displayed twice. But in the Google result information that the result also retrieved from Yahoo will be added.

4) Categorizing and displaying results

Results from the filtering process will be the non-redundant results. For displaying purposes the filtered results will be populated in new array in the form that it can be categorized by the source of retrieval. Let say the result was being retrieved from Google, Yahoo and MSN, this group will be displayed on the top list followed by group retrieved from Google and Yahoo, Google and MSN, Yahoo and MSN, Google only, Yahoo only and last but not least MSN only.

Stage 5: System Testing

System testing occurred not only after system production stage but it occurred throughout the development stage to ensure every stage completed as what was required. The testing was done with the main goal to assess the extent of the effectiveness of the system to fulfill the user's need. The effectiveness of the system was measured by how far this system helps users in saving time searching information on various search engines. It was also measured by how effective this system presents the results to the users and how well the ranking was done according to the users' needs and convenience.

This system was developed part by part. It has three major units which are queryGoogle unit, queryYahoo unit, queryMSN unit. Each unit was being tested separately for its functionality and effectiveness in populating the results retrieved.

System testing was done after each of unit testing was thoroughly done. After combining all the three units the combined system need to be combined for the compatibility and flow of the functions. During the system testing, it took sometimes for the system to function as what it should be as the units are not compatible with each other. To test the system the unit need to be fully working and each of the unit should be able to interact with each other. The first part of the system testing is to test the system functionality. This is to test whether all results from all search engines can be combined and populated as one search results.

After this stage it is then to test the filtering process effectiveness. How effective this search engine filters out any identical URL. The comparison of the total results retrieved from the Google, Yahoo! Search and MSN Search and total results from IFWSE after being filtered was calculated. Precision theory being used to calculate the preciseness of the filtering process of this search engine.

3.2 Tools Used

The specification of the hardware used to develop this system and the minimum requirement of hardware to run the system are as follow:

Developing Hardware	Running Hardware Requirement
PC Pentium4 1.5 MHz	PC 400 Pentium II
640Mb RAM	32Mb RAM
64Mb Graphic Memory	4Mb Graphic Memory
20Gig HDD	6Gig HDD

Table 3.0: Hardware Requirement

Development tools used in this project are as follows:

- Macromedia Dreamweaver MX 2004 v7.0.1

This tool is used as an aid to develop the interface. Suitable to the web-based nature of the system, it was developed on PHP scripting language.

- Adobe Photoshop and Adobe Image Ready 7.0

The developer uses these two tools to design and manipulate graphics and images. It provides all sorts of alteration tools, to enhance the system's interface. The main purpose the developer use this tool was basically to design logo and interactive and attractive fonts.

- EasyPHP

A tool used by the developer to ease the PHP, Apache and MySQL installation and configuration. It is a 3 in 1 tool that enables the developer to install all those three with only one installation. In this version of EasyPHP, it has Apache 1.3.27, PHP 4.3.3, MySQL 4.0.15 and phpMyAdmin 2.5.3.

PHP scripting language was being used as the language programming to develop the interface. For this the PHP package was needed in order to compile the coding. MySQL was the database used while developing the system prototype. The implementation of MySQL was aided by phpMyAdmin, the GUI version of MySQL. It eased the creation of tables in MySQL.

- Microsoft Visio

This tool was used mostly during the design phase where the writer developed the diagram such as Process Flow Diagram and System Architecture illustration.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Results

Table 4.0 and 4.1 both show the results achieved after performing number of experiments with IFWSE to acquire the total IFWSE results retrieved. 20 different keywords are being tested and the total results retrieved from each and every search engine are being recorded to be compared with the filtered results from IFWSE.

Keyword Searched	Retrieve from	Total Results						Algorithm
		Google	Yahoo! Search	MSN Search	Total Results from all 3	FWSE	Meta Search Engine	
Marketing Mix		50	50	50	150	104	129	138
Precision and Recall		50	50	50	147	136	108	129
Intelligent Interface		50	47	50	152	147	148	149
Petrobras		50	50	50	150	135	139	138
Brazil		50	50	50	151	148	149	150
Oil and Gas Plant in		50	50	50	148	149	149	150
Ontology Based Application		50	50	50	149	149	149	150
Google Page Rank		50	50	50	149	149	149	150
Page Hits Ranking		50	50	50	149	149	149	150
Results		50	50	50	149	149	149	150
Meta Search Engine		50	50	50	149	149	149	150

Table 4.0: Comparison total results retrieved-1

Keyword Searched		Total Results					
		Retrieval from					
		Google	Yahoo! Search	MSN Search	Total Results from all 3	IFWSE	Nano Technology
Digital Divide		50	51	50	150	151	115
Data Mining		50	50	51	151	108	108
Geographical		50	50	50	150	132	118
Information System		50	50	50	150	139	137
Waterfall Process		50	50	50	150	134	148
Open Source		50	50	50	150	151	151
Software		50	50	50	150	154	135
Define Process		50	50	51	151	137	137
before output in		50	50	50	150	148	148
ABAP		50	50	50	150	151	151
Download		50	50	51	151	154	154
Multimedia audio		50	50	50	150	151	151
Controller device		50	50	50	150	151	151
Free Hosting Server		50	50	50	150	151	151
Nano Technology		50	50	50	150	151	151

Table 4.1: Comparison total results retrieved-2

Total results retrieved from each of the search engine would be around 45-55 as IFWSE retrieved the results from first 5 pages. Compare to the total results for all three search engine, total search results by IFWSE is lesser by about 1/4.

Let's take the first searched keyword 'petronas' as an example. The total results from three search engine are hundred and fifty but the IFWSE result is only ninety-nine. In details, overlapped results for all three **Google-Yahoo! Search-MSN Search** were seventeen, **Google-Yahoo! Search** were ten, **Google-MSN** were six, **Yahoo! Search-MSN** got only one overlapped result and the rest of the results were unique and non-redundant with the other results. From the observation, results that overlapped normally high ranked results that retrieved from 1st top and 2nd top pages. The remaining pages seldom got overlapped results.

4.2 Discussion

Looking at the results retrieved, Integrated Filtered Web Search Engine obviously can filter the redundant results out of each of the search engines. This is done so that, later on each of the results that have identical URL be displayed only once for users benefits.

Based on the results we can say that for only top five result pages of the search engines $\frac{3}{4}$ of the results are unique save about $\frac{1}{4}$ of the time from reading the similar results. Here is the evidence where search engines' results overlap far less than we would think. That's the reason why users constantly have the habits of opening more than one browser for another search engine.

The objective of IFWSE included filter up results to eliminate any redundancy as well as to give users more top ranking results taken from many search engine. The successful of the first mentioned objective is being evaluated using Precision theory where it stresses on how high the precision of the filtering process is. And the success of the second mentioned objective is being evaluated by the comparison of total results returned by IFWSE and total results returned by MetaCrawler (Meta search engine).

4.2.1 Precision

Table 4.2 shows the preciseness of this IFWSE filtering process. Number of redundant results got from a very close inspection of the results during the experiments.

Keyword Searched	Total results retrieved		
	IFWSE	Redundant results	Precision (%)
Petronas	99	4	96%
intelligent	121	2	98%
Intelligent interface	131	3	98%
Precision and recall	120	1	99%
Marketing mix	113	4	96%
Oil and gas plant in brazil	135	3	98%
Ontology based application	141	4	97%
Google page rank	109	1	99%
Page hits based ranking results	129	1	99%
Meta search engine algorithm	138	1	99%
Digital Divide	115	2	98%
Data Mining	108	1	99%
Knowledge Management	118	3	97%
Geographical Information System	132	5	96%
Waterfall Process Model	139	5	96%
Open Source Software	117	3	97%
Define process before output in ABAP	134	4	97%
Download multimedia audio controller device driver	148	7	95%
Free hosting server	137	6	96%
Nano technology	135	3	98%
Average	125.95	3.35	97%

Table 4.2: Precision Table

Some inspection was done to each of the results retrieved and it is found that this IFWSE still got approximately up to five redundant results after being all filtered. Based on this we can calculate the precision of the filtering process by using the following formula.

$$\boxed{\text{Precision} = (A / B) \times 100\%}$$

A = number of non-redundant (total results returned – number of redundant results)

B = total results returned

E.g.

Searched keyword: petronas

Total results returned: 99

Redundant results: 4

$$\boxed{\begin{aligned}\text{Precision} &= (95/99) * 100\% \\ &= 96\%\end{aligned}}$$

The precision was calculated for each of the searched keyword and average of precision was calculated. Basically IFWSE can approximately filter up the results with 97% precision which means around 2-3 results are redundant. These redundant trends are actually those results from Yahoo! Search and MSN Search. It is believed that, the performance of the filtering process for the second round becomes less effective. This is due to the structure of the algorithm

4.2.2 IFWSE vs. MetaCrawler

Table 4.3 shows the comparison of total IFWSE results to total MetaCrawler results.

Keyword Searched	Total results retrieved	
	IFWSE	MetaCrawler
Petronas	99	60
intelligent	121	65
Intelligent interface	131	82
Precision and recall	120	72
Marketing mix	113	82
Oil and gas plant in brazil	135	96
Ontology based application	141	80
Google page rank	109	76
Page hits based ranking results	129	63
Meta search engine algorithm	138	78
Digital Divide	115	81
Data Mining	108	81
Knowledge Sharing	118	106
Geographical Information System	132	91
Waterfall Process Model	139	87
Open Source Software	117	97
Define process before output in ABAP	134	68
Download multimedia audio controller device driver	148	93
Free hosting server	137	87
Nano technology	135	104

Table 4.3: Comparison of total result of IFWSE and MetaCrawler

Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. Basically it is used to show the percentage of desired results and successfully retrieved. It is difficult to measure recall as WWW is such a huge warehouse and we don't know how many relevant results in it [J.Richard-2000]. So instead of using recall to measure percentage of relevant results retrieved, comparison between IFWSE and MetaCrawler is used.

IFWSE uses only three search engines compare to MetaCrawler that uses up to five search engines. Still numbers of retrieved results by MetaCrawler are much smaller compare to IFWSE. Rational behind this is because results retrieved by MetaCrawler got more overlapping results compare to the results retrieved by IFWSE. Why is this happening? Based on the close observation done, MetaCrawler retrieve up to top 3 top pages of the results in each search engine whereas IFWSE retrieves up to top 5 pages from each of the search engines.

With top 3 search pages for 5 search engines MetaCrawler can get up to 150 results without filtering and after filtering process this total results returned would go nearly half of the unfiltered total results. With IFWSE that should also get 150 results without filtering process, normally got up to $\frac{3}{4}$ of the total unfiltered results. To say that, IFWSE got some redundant results, this might be true but this problem contribute to only 0.5 percent of the larger total results returned.

From the observation done during the experiments, most of the overlapped results are the high ranked. The lower the ranked the more unique the results returned. So we could say that, the results retrieved from MetaCrawler are mostly at the top 30 and mostly these results highly overlapped with the other results from other search engines. Whereas IFWSE retrieved the top 50 results that are less overlapped with each other. So basically, IFWSE got the advantage or variety results returned to the users.

CHAPTER 5

CONCLUSION AND RECOMMENDATION

5.1 Conclusion

Search engine is one of the popular tools nowadays in information searching on the web. Because searching information on the web is not easy and user needs to have some tools to aid them in finding the desired information. We have Internet Directories such as Yahoo! Directory, Search Engines that crawl the web such as Google, Yahoo! Search, and also Meta-Search that retrieved information from other search engines simultaneously such as Meta-Crawler, Dogpile and etc.

“These days, we can find more than ever, faster than dreamed of, but we’re also taking it for granted. Information at your fingertips; when you have a question, fire up Google. The answer’s out there.”

-Philippe Lenssen- [S. Chris-2005]

But still, human is very hard to please. Even though the search engine is sophisticated beyond their imagination still they have inconvenience where their ease in use matters. Users have the tendencies to use more than one search engine to get a better results and because they have this urge to know what actually other search engine get that their search engine do not get. Nowadays, we even have the intelligent search agent like Copernic that can do the searching intelligently but it needs to be installed before used. And there also Meta-search engine that can simultaneously can search for various search engines and present the results in one page. But those were not filtered. But as for today there are some meta-search engines

that filter their retrieved results but still major search engine like Google, Yahoo! Search and MSN Search is not.

Basically Integrated Filtered Web-Search Engine' objective is to integrate all the major search engine which are Google, Yahoo! Search and MSN Search into one filtered search engine for user convenience. It will benefit the users in saving their time reading results from several opened browsers. Users also can get more results with one single search as the IIWSE will simultaneously send queries to several search engines at once. Other than that, it will filter out any redundant results and display them only once so that users do not have to read them twice. This will give the users benefits over the unfiltered meta-search engine.

Last but not least this Integrated Filtered Web-Search Engine (IFWSE) **combines the power** of the top three most popular search engines.

5.2 Recommendation

For future enhancement it is recommended that this IFWSE add more major search engines for more retrieved results. As for the filtering process, instead of only filtering only base on the identical URL to make it more effective, results refinement method can be used such as Boolean search. This enable the Boolean search refinement methods applied by remote search engine be applied in IFWSE as well for higher quality results. Other than that, the filtering algorithm could be improved to make it more precise.

REFERENCES

[B. Sergey and P. Lawrence] B. Sergey and P. Lawrence, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", *Computer Science Department, Stanford University, Stanford, CA 94305*

Available at:

<http://www-db.stanford.edu/~backrub/google.html>

[Retrieved on April 26, 2006]

[B. Joe – 2005] B.Joe, 23 August 2005, "Meta-Search Engines", *Copyright (C) 2005 by the Regents of the University of California*

Available at:

<http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/MetaSearch.html>

[Retrieved on April 28, 2006]

[Copernic-2005] Copernic, about us -- Technologies, *Indexing*. October, 2005.

Available at:

<http://www.copernic.com/en/company/technologies.html>

[Retrieved on April 02, 2006]

[G. Michael Youngblood-1999] G. Michael Youngblood, 1999, "Web-Hunter: Design of a Simple Intelligent Web Search Agent", *CSE Department, University of Texas, Arlington*

Available at:

<http://www.acm.org/crossroads/xrds5-4/webhunting.html>

[Retrieved on April 02, 2006]

[G. Robyn- November 14, 2002] G. Robyn, November 14, 2002, "Search Engine Usage Ranks High", *Copyright 2006 Jupitermedia Corporation All Rights Reserved*. Retrieved Available at:

<http://jbr.org/articles.html>

[Retrieved on April 02, 2006]

[J. James-1996] J. James, 1996, "Using Intelligent Agent to Enhance Search Engine Performance", *First Monday*, the Peer-Reviewed Journal on the Internet Available at:

http://www.firstmonday.org/issues/issue2_3/jansen/

[Retrieved on April 26, 2006]

[J.Richard-2000], J.Richard, 2000 "Measuring Search Effectiveness", *Creighton University Health Sciences Library and Learning Resources Center*

[Liu-1998] Liu, Jian, "Guide to Meta-Search Engines", *BF Bulletin (Special Libraries Association Business and Finance Division)*. 107 (Winter 1998): 17-20.

[Neal-1997] Neal Harper, 1997, "Intelligent Agents and the Internet", *COM336 Artificial Intelligence*, University of Sunderland, School of Computing

Available at:

<http://oasis.sunderland.ac.uk/cbowww/AI/TEXT/AGENTS3/agents.htm>

[Retrieved on April 26, 2006]

[NLP] Natural Language Processing in Information Retrieval,

Available at:

<http://www.searchtools.com/info/ir-nlp.html>

[Retrieved on October, 2005]

[N. Masoud-2003] N. Masoud, Fall 2003, “Web Intelligence Conceptual Search Engine and Navigation”, *BISC Program, Computer Sciences Division, EECS Department University of California, Berkeley, CA 94720, USA*

[Regina-2004] Regina Hayati Rahiman, December 2004 *Integration Tool to Integrate Popular Search Engine as One Main Search Engine*, Bachelor Degree Thesis, University Technology PETRONAS, Business Information System.

[R.I. John, G.J. Mooney-2001] R.I. John, G.J. Mooney, “Fuzzy User Modeling for Information Retrieval on the World Wide Web”, *Knowledge and Information Systems* (2001) 3: 81-95

[S. Chris-2005] S. Chris, 2005, “Internet Search Strategies: Search Tools”. *Minnesota West Home*, Minnesota West Community & Technical College
Available at:

<http://www.mwetc.cc.mn.us/libraries/strategies/tools.htm>

[Retrieved on April 26, 2006]

[S. Danny- July 31, 2003] S. Danny- July 31, 2003, “How Search Engines Rank Web Pages”, *Search Engine Watch*
Available at:

<http://searchenginewatch.com/webmasters/article.php/2167961>

[Retrieved on April 26, 2006]

[S. Danny- January 24, 2006] S. Danny, January 24, 2006, “Nielsen NetRatings Search Engine Ratings”, *Search Engine Watch*

Available at:

<http://www.searchenginewatch.com/>

[Retrieved on April 26, 2006]

[S. Fabrizio-2004] S. Fabrizio, May 2004 *High Performance Issues in Web Search Engines: Algorithms and Technique*, Ph.D. Thesis, UNIVERSIT `A DEGLI STUDI DI PISA.

[S. Vrettos, A. Stafylopatis-2001] S. Vrettos, A. Stafylopatis, “A Fuzzy Rule-Based Agent for Web Retrieval Filtering”, *N. Zhong et al. (Eds.): WI 2001, LNAI 2198, pp. 448-453, 2001.*

[T.M.T. Sembok-2003] T.M.T. Sembok, “Character Strings to Natural Language Processing in Information Retrieval”, *et al. (Eds.): ICADL 2003, LNCS 2911, pp. 26-33, 2003.*

APPENDICES

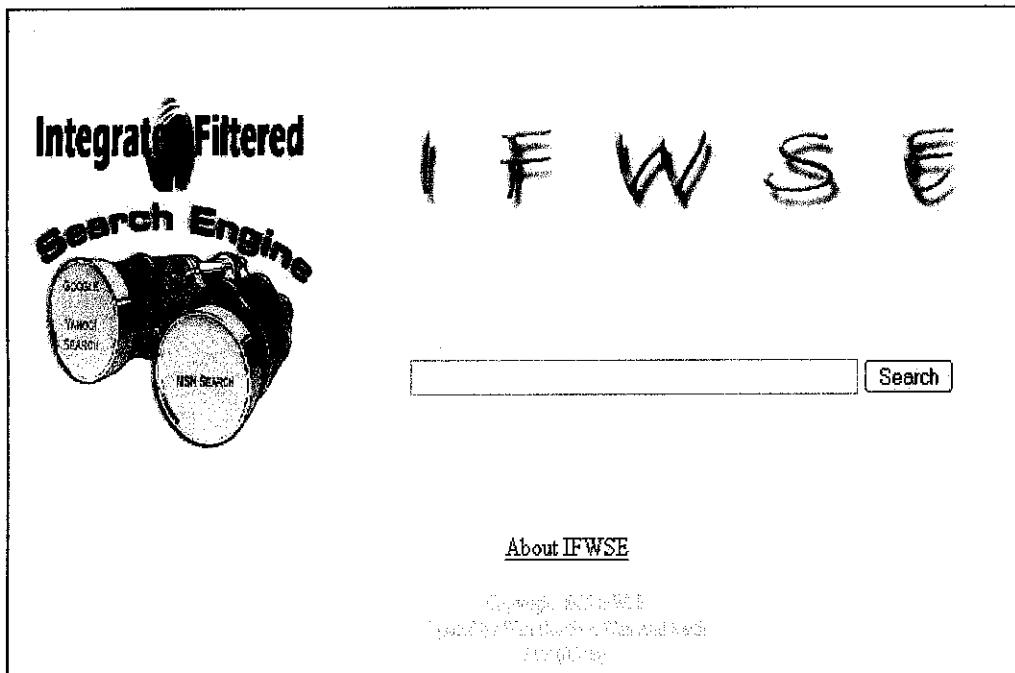


Table A: Index Page

Searched Keyword:	Total Results:	From
digital divide	119	
<hr/>		
Equity Digital Divide Campaign		
<p>The Digital Divide. In essence, the digital divide is the difference in access to learning resources that modern technology offers young people, usually a working computer and an Internet connection.</p>		
<p>MSN</p>		
<ul style="list-style-type: none">♦ www.equitycampaign.com♦ Cached page		
<hr/>		
AIBD: Publication on Responses to Globalization and the Digital Divide ...		
<p>Responses to Globalization and the Digital Divide in the Asia-Pacific: The 1st Conference of the Ministers on Information & Broadcasting in the Asia Pacific Region, May 27-28 2003, Bangkok, Thailand</p>		
<p>MSN</p>		
<ul style="list-style-type: none">♦ www.aibd.org.my/page/www_publications/books/42.html Cached page 6/17/2006		

Table B: Result Page