DEVELOPMENT OF AUTOMATED WEB TRAVERSING TOOL

TAI SOCK YIN

FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2004

DEVELOPMENT OF AUTOMATED WEB TRAVERSING TOOL

TAI SOCK YIN

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF COMPUTER SCIENCE

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY MALAYA
KUALA LUMPUR

2004

Abstract

As the size of World Wide Web (WWW) grows rapidly and relevant web sites proliferate, the issue of locating information becomes increasingly challenging. We, in Malaysia are among the 215 million Internet users within the South East Asia region (SIL, 2000), who also show an exponential growth in numbers of web pages, similar to the trends of WWW in general. Thus, collecting Malaysia pages becomes a tough problem. To manually check out pages from some possible portals, directories or even search engines require considerable amount of time and effort. A significant aspect of finding these pages is the set of choices for automatically traversing from one web page to another and the ramifications that these choices have will provide different search results. This study investigates the development of an automated traversing prototype that implements breadth first and depth first approaches to gather Malaysia web pages from the WWW, which will allow the organized study of the navigational aspects of web site. Finally, it describes how the use of these traversal approaches can achieve different results. The dissertation therefore involves work that spans in three major areas. First, understand the structure of the web as a directed but unstructured graph, as well as familiarize with the two elementary traversing approaches. Secondly is to build a working prototype of traversing tool to experiment the traversing approaches. Finally, is to investigate on how to examine the quality of web pages gathered by two different traversing approaches, in terms of two aspects: recall (measure of the ability of the prototype to find all of the relevant items that are in the database) and precision (a measure of accuracy of the traversing process).

Acknowledgement

I would like to express my gratitude to all the people who have helped me in completing this dissertation.

First, I would like to thank my supervisor, Associate Professor Dr. Zaitun Abu Bakar, for her encouragement, guidance and assistance through my graduate studies.

I am greatly thankful to some of my colleagues for their helpful comments. I would like to thank particularly Mun Soon, Siti Salleh and Seow Hoon.

Special thanks to Mr. Robinson a/l J. Samual for proof reading my dissertation.

Last but not least, I dedicate this dissertation to all my family members who always support me and listen patiently to me.

Table of Content

LIST OF FIGURES

LIST OF TABLES

1.0 INTRODUCTION

1.1 Background and Motivation

Today's Internet has multiple usages, ranging from electronic commerce to online education. To many of us, the Internet is not more than a place to find some information (Kowalski and Maybury, 2000). Hence, it is reasonable to consider the ultimate purpose of the Internet remains in providing information. This can also be traced back to its predecessor, the ARPANET, where the dominating development factor was to support information sharing and exchange, mainly for government and academic researchers (William, 2000).

In spite of FTP (File Transfer Protocol), telnet, emails, and Word Wide Web (WWW), the widespread use of different mechanisms for searching items is also a visible development of the Internet phenomenon. The primary approaches referred are usually associated with search engines or online directories on the Internet. They create indexes of items and offer user-friendly interfaces with search function as the most basic and efficient way to find useful information. Some of the most commonly used search engines are YAHOO, Alta Vista, Lycos (Kowalski and Maybury, 2000) as well as some newer alternatives such as Google, FAST Search etc. In Malaysia, we have domestic search engines, for example CARI.com, Catcha.com, and Lycos Malaysia, which are specifically customized to suit the local information needs.

Closely related to search engines but transparent to the users are the traversing processes that effectively gather all the fast-growing information. Usually, the information discovered is catalogued and subsequently accessed by users through a GUI

(Graphical User Interface) that provides at least a "search button". By clicking on the button, users are hoping that the information required will be delivered to their monitors.

However, with the volume of information on the Internet growing exponentially to an astronomical figure, the process of finding relevant information becomes a painful experience to many users. In response to this tedious and lengthy Internet search, traversing processes are no longer limited to search engines to populate their databases at the backend, but are directly available to users via online information systems (Kowalski and Maybury, 2000).

1.2 Problem Statement

In general, the process of searching the Internet usually starts with either visiting large number of the existing sites or retrieving web pages that have been readily indexed by search engine. Such simple approach requires users to follow the hypertext links one by one to view items. Unfortunately, due to the huge size of the Internet, manually navigating through the hyperlink hierarchy by expanding the link on a particular topic seems less practical at the current time. Furthermore, the International Data Corporation (Gantz and Glasheen, 1999) estimates that there are 3 millions new web pages or 59 gigabytes of text being created every day and it is estimated that the number of web pages will exceed 16.5 billion by 2003. The loose structure of the overall connectivity allows individuals and organizations to become part of the Internet and publish information easily. This clearly shows that the whole world is slowly but surely getting connected. Such endless expansion continues to generate large amount of information that certainly overwhelm anyone performing a search on the Internet.

This overwhelming information has introduced both opportunities and threats to the Internet. As a result of the difficulties to subdue the information overload problem, many new and potential research development areas are booming, including technologies to automate the human process to define items of interest and go to various sites searching for the desired information (Kowalski and Maybury, 2000).

Despite the fact that information overload is normally perceived as a global issue, there are also new and unique challenges for exploration in more local context. It is reported that Internet users in the Asia-Pacific are continuing to show a clear preference for viewing local web pages (Ong et al., 2000). Malaysians are among the 215 million Internet users within the South East Asia region (SIL, 2000), which relates to the exponential growth in numbers of web pages, similar to the trends exhibited by the WWW in general. With Malaysia gearing towards the knowledge-based economy, more Malaysians are getting online and providing knowledge content on the Internet. This increases the needs of both Malaysians and foreigners in finding and retrieving important national information from the web.

Malaysia web pages here refer to any web page on the Internet that discusses about Malaysia in any aspect including social, economy, politics, geography, history, business, news etc. (Please refer to section 4.2.4, which provides a more detailed definition of Malaysia web pages). These pages are usually scattered across the web. Some can be easily identified from their web address, for example those ended by ".my", but there are also some pages that do not use ".my" address. Another way would be to check out pages from some possible portals, directories or even search engines. But this necessitates substantial amount of time and effort. Moreover, as the number of

web pages continue to grow, manually surfing the Net to find these pages seems less feasible in the long term. Thus, collecting Malaysia pages becomes an increasingly interesting but "tricky" problem. There are at least two major technical challenges involved here. First, the web is an unstructured graph without any guidelines for searching direction. Second, there is a high volume of web pages available online that is impossible to collect manually. Hence, the solution requires research into web structure, implementation of traversal strategy and development of an automated system.

1.3 Research Objectives

The main objective of this dissertation is to investigate the development of an automated traversing prototype that implements breadth-first and depth-first approaches to gather Malaysia web pages from the WWW.

The objective therefore involves work that spans in three major areas. First, understand the structure of the web as a directed but unstructured graph, as well as familiarize with the two elementary traversing approaches. Secondly is to build a working prototype of traversing tool to experiment the traversing approaches. Finally, is to investigate on how to examine the quality of web pages gathered by two different traversing approaches, in terms of two aspects: recall (measure of the ability of the prototype to find all of the relevant items that are in the database) and precision (a measure of accuracy of the traversing process).

1.3.1 Scope

The main challenge in this work lies in how to automatically and independently move between web sites locating web pages from the Internet. There are several different

ways to automatically traverse the Internet. Among all, breadth-first and depth-first are the most fundamental approach (Barforoush et al., 2002). Therefore, they were selected for this preliminary investigation. The information available on the Internet is made up of text and non-text format such as image, audio and video but this dissertation is only focusing on locating textual information.

1.3.2 Contribution

Locating information on the Internet can be as difficult as finding a needle in a haystack. The ultimate expectation of this dissertation is to contribute towards the effort to subdue the difficulty of finding relevant information on the web, more specifically Malaysia web pages. Apart from that, the analysis result, which focuses on local web pages as data set, reveals some Malaysia web pages development trend, viewing from the context of web structure.

1.4 Research Questions

The main research questions to be addressed in this dissertation include the following:

- What is the structure of the hypertext linkages of the web?

  The size of the web changes every day, hour, minute and second. The vastness of the information space has given rise to questions on how the structure of hypertext links is organized and arranged. Consequently, how to make use of the existing web structure to find Malaysia web pages?

- How to apply the conventional traversing approaches to traverse the cyberspace?

  Imagine a sea of information, and a 'sailor' with an unknown destiny, in search of an answer. Without proper guidance, the sailor would surely be lost. The web

is very much like this sea, and many users are like the sailor. Implementing basic approaches (i.e. breadth-first and depth-first) for traversing the web also involves work on how to design and perform heuristic searching plan on the structure identified. The heuristic for deciding which node to traversal next follows the page-based approach (more description in section 3.2, page 56).

- What is the recall and precision percentage of textual information found using the approaches above?

  A more complex area of measurement for any information system is connected with the search process (Kowalski and Maybury, 2000). The recall and precision of information system evaluation measures results of the breadth-first and depth-first traversing approaches.

1.5 Research Methodology

The research methodology selected to complete the investigation is quantitative study through experiment. The following steps were carried out to produce the research outcomes and Figure 1.1 on page 7 summarizes the steps and output of research methodology used.

- Form hypothesis on the web structure and traversing approaches by reviewing existing literature.
- Develop working prototype to implement the traversing approaches.
- Design traversing plan to collect data focusing on Malaysia web pages.
- Experiment by running the prototype.
- Analyze experimental results using the precision and recall as measurement.
- Draw conclusion and recommendations based on the findings.

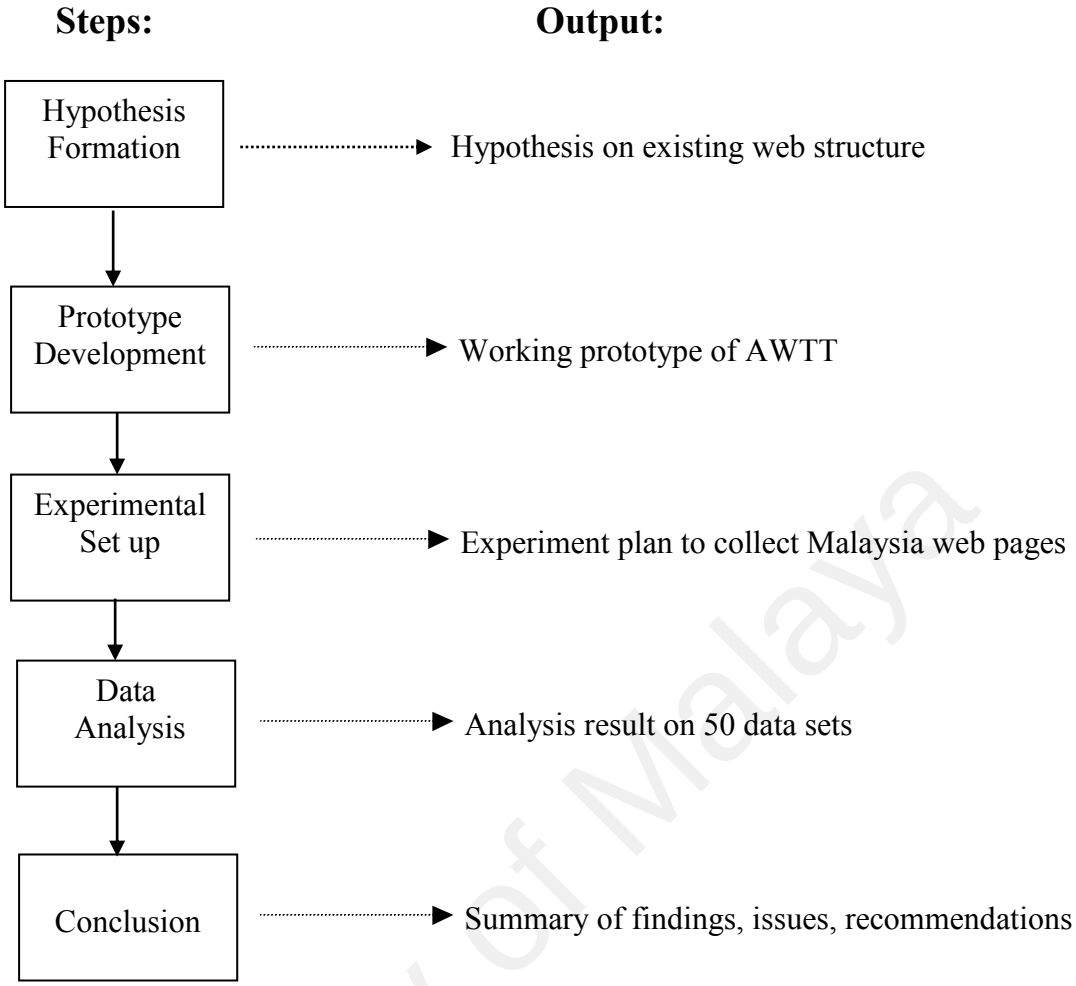


Figure 1.1
Summary of Research Methodology

1.6 Expected Research Outcomes

Below is the summary of expected research outcomes of this dissertation:

* Prototype of Automated Web Traversing Tool (AWTT)

  Development modules:

  - Traversing approaches: breadth-first and depth-first

  - Evaluation measurements: recall and precision

  - Database

  - GUI

* 10,000 web pages

* Experimental result

  50 collections of different experimental datasets per traversing approach

1.7 Organization of Dissertation

Below is the outline of the dissertation:

Chapter 1.0 is the introductory chapter to this dissertation. This chapter explains the background and motivation, problem statement, research objectives, scope, contribution, research questions, research methodology, expected outcomes and the organization of dissertation.

Chapter 2.0 reviews some compelling findings or research breakthrough in three areas: the web structure, traversing approaches and information systems evaluation. In the first section, the subject areas on the research of web structure, which are closely related to the dissertation are examined and highlighted. It also differentiates the web search with conventional search to take note of the issues relevant to automated traversing prototype development and experimental set up. Several more popular traversing schemes or approaches are analyzed and compared in the second section. The third section studies the possible use of the existing information systems evaluation approaches for measuring the search results.

Chapter 3.0 provides an overview on the development framework and architecture of the automated traversing prototype. Major components of the working model are identified and the basic functionalities as well as integration are described.

Chapter 4.0 elaborates the data preparation process, the steps involved throughout the experiment and the problems encountered during the experiment. This chapter gives some significant observations through data analysis and some insights generated from the experimental data obtained in previous section.

Chapter 5.0 discusses the data analysis and experimental results in more detail based on the precision and recall measurement, and then compare it to some earlier work as described in the literature review.

Chapter 6.0 concludes the dissertation with summary of the work. It highlights some limitations of the current system and also includes several suggestions for improvement. In addition, it recommends some potential areas for further research with respect to the current trends and directions.

Appendixes enclosed detailed experimental reports, source codes and some sample raw pages collected.

2.0 LITERATURE REVIEW

2.1 Exploring the World Wide Web (web)

This section explores the web hypertext links structure based on some recent findings. The concepts reviewed were used to form hypothesis for certain aspects of this investigation. More importantly, it answers the question of how to utilize the existing web structure to find pages, which discuss on Malaysia (as stated in section 1.4). By having a clearer picture of how web pages are connected as well as understand some important characteristics of the web, a prototype can be built and more comprehensive traversing strategy can be planned to facilitate the process for collecting Malaysia web pages.

2.1.1 Difference between the Internet and the Web

Most Net users naturally see the Internet and the web as one and the terms are used interchangeably. Conceptually, there is a minor difference between the two technologies as pointed out by Vincent (1995). The relationship between the Internet and the web is analogous to a car and its engine. For example, the EON engine can be wrapped into different models such as Proton Wira or Proton Iswara. Similarly, the Internet is the core engine that provides the connectivity throughout the world but it is accessible via other application technologies. World Wide Web, FTP, telnet, and emails are "wrapper" that users are more familiar with. And among all, the web happens to be the most popular "wrapper" because of the frequent use of web sites (Gantz and Glasheen, 1999). Unsurprisingly, the Internet is always referred to the World Wide Web because the web is essentially seen as a great looking body wrapped around a high-performance engine –

the Internet. Using the car analogy, the web can be used to drive and control the Internet, but the Internet is still providing all the horsepower (Vincent, 1995).

This discrepancy deserves a special attention to avoid digressing beyond the scope of this dissertation when investigating the development of traversing tool in later section, whereby focus is placed in the web pages rather than other applications, i.e. FTP, telnet, email hosts and servers etc.

2.1.2 Self-Organization of the Web

Due to the freedom, simplicity and convenience of publishing information online, the web is widely perceived as having lack of structure and it is unmanageable. However, some recent studies (Kumar et. al., 1998, Barabasi and Albert, 1999, Broder et. al., 2000, Albert et. al, 1999) that are based on the experiments on local and global properties of the web graph have shown a great deal of self-organization.

In the global perspectives, the distribution of hyperlinks among pages is taken into consideration for analysis. Studies above show that the number of links to and from individual pages is distributed according to the power law over many orders of magnitude; fraction of pages with n in-links is roughly $n^{-\alpha}$ for $\alpha \sim 2.1$. This means that as the web grows by the sequential arrival of new web sites, the probability that an existing site gains a link is proportional to the number of links it currently has (Kleinberg, 1998). In this case, when web pages are randomly added, the web tends to organize itself in a "rich-get-richer" process, where pages with more existing links continues to enjoy higher chances in linking with other pages on the web. Nevertheless,

the study results are still insufficient towards a complete understanding of the governing process for web self-growth, because some obvious deviations appear when local perspectives are examined.

The studies also involve the local level analysis, which is focused on smaller scale of neighbourhoods and regions, for example university homepages, company homepages or Asian homepages etc. The analysis results present two ways of viewing the self-organization structure. However, both ways also show that pages and links are created by users with particular interest and pages on the same topic tend to cluster into natural "community" structures that exhibit an increased amount of links (Kleinberg and Lawrence, 2001).

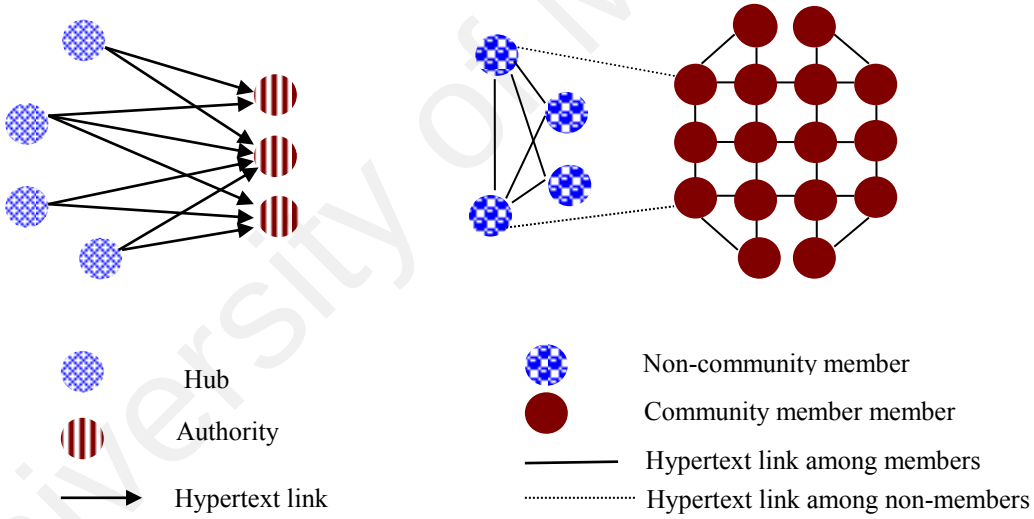


Figure 2.1.1(a) Figure 2.1.1(b)

Hubs and Authorities Organization of Web Pages
(Kleinberg and Lawrence, 2001)

The first method reveals that a small set of pages with extraordinarily large number of links may be a sign of topically related pages. On the other hand, the second method points out that a "community" can also be represented by a collection of pages in which

each member page has more links to pages within the "community" than to pages outside the "community".

As described, the web actually demonstrates a remarkable degree of self-organization in the link structure and clusters the web pages into related subjects. As depicted in Figure 2.1.1(a), web pages can be defined as hubs and authorities. A hub is a well-known web page that contains links pointing to many authorities (less famous web pages), whereas an authority is a web page that is pointed by many hubs (well-known web pages).

Besides, the web can be alternatively seen to have arranged into "communities" based on number of links that is greater among members than between the rests of the pages in other groups as shown in Figure 2.1.1(b).

The traversing strategy developed in later section is built on the same notion above, whereby web pages discussing on Malaysia is perceived to have naturally organized into a "community" which is close to each other in the link structure.

2.1.3 The Web Structure Based on Hyperlinks Distribution

To common knowledge, the web never stops expanding since its creation and has transformed into a vast information repository. Unlike many well-established networking systems in the country - the railway and highway systems (Kleinberg and Lawrence, 2001), such as KTM, STAR LRT, Putra Link, Federal Highway, Kesas Highway etc); the web is a self-organizing system without proper designed and planned architecture for continuous growth. Rather, it is a virtual network of content and

hyperlinks, with over a billion interlinked "pages" generated by the uncoordinated actions of ten of millions of individuals (Kleinberg and Lawrence, 2001). The decentralized, unpredictable and dynamic nature has caused tremendous challenge in determining its structure. Recent research results obtained range from theoretical (e.g. models for the graph, semi-external algorithms), to experimental insights (e.g. regarding the rate of change of pages, new data on the distribution of degrees) and practical (e.g. improvements in traversing technology) aspects. This has disclosed the mysterious face of the web in certain extends (Broder, 2000).

Andrei Broder (2003), an IBM distinguished engineer and the CTO (Chief Technology Officer) of the Institute for Search and Text Analysis in IBM Research, highlighted the increasing interest in web structure recently. As addressed in his keynote speech for SIGIR (Special Interest Group in Information Retrieval) 2003 Conference, the web graph, meaning the graph induced by web pages as nodes and their hyperlinks as directed edges, has become a fascinating object of study for many people including physicists, sociologists, mathematicians, computer scientists, and information retrieval specialists. This area of research provides valuable idea for the work here, particularly on how to design and perform a heuristic searching plan for the traversing tool to move across the web as implemented in chapter 4.0.

A study conducted by A. Broder et al. (2000), which is based on the experiments on local and global properties of the web graph using two Altavista crawls each with over 200 million pages and 1.5 billion links produces an interesting macroscopic view of the

structure of the web. This microscopic view reduces the abstraction of web and provides an image that is visible to net users' naked eyes.

The study considers the web as a directed graph whose nodes correspond to static pages, and whose arcs correspond to hyperlinks between these pages (as shown in Figure 2.1.2 (a)); various properties of the web including its diameter, degree distributions, connected components, and macroscopic structure has been investigated during the study.

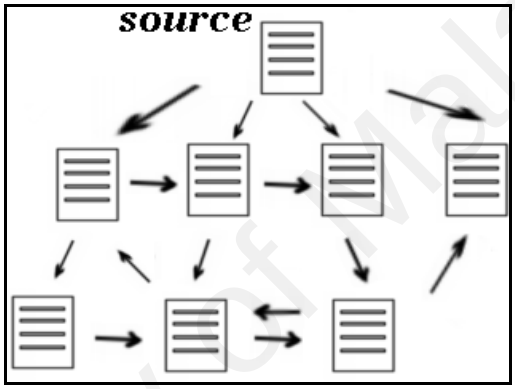


Figure 2.1.2(a)
A Directed Graph with Web Pages as Nodes

The study was presented in the 9th Web Conference for the first time (A. Broder et al., 2000). This experimental evidence reveals a rather more detailed and subtle picture of the web. Results conclude that significant portions of the web cannot at all be reached from other significant portions, and there is significant number of pairs that can be bridged, but only using paths going through hundreds of intermediate pages. In other words, the result of the study indicates that the web contains large, strongly connected web pages, in which every page might not reach every other directly but can be connected by a path of hyperlinks.

As the study drilled down to a deeper level, more interesting facts were observed. The researchers found that all the web pages can be divided into two categories: core and non-core sites. The first set of web pages contains most of the prominent sites on the web and is commonly known as the "core". Whereas, the remaining non core pages, characterized by their relation to the core, can then be further divided into another three smaller categories: upstream, downstream and "tendrils".

As depicted by IBM Almaden Research Lab (2000), Figure 2.1.2(b) illustrates the macroscopic view of the web constructed based on the newly emerged theory generated from the findings above, known as Bow Tie Theory. This theory explains the dynamic behaviour of the web, and yielded insights into the complex organization of the web. The theory emphasizes that the image of the web looks very similar to a bow tie, made up of four distinct regions described earlier - the core, upstream, downstream and the tendrils. The remaining discussion of this section is based on the findings reported by Almaden Research on how to map the web.

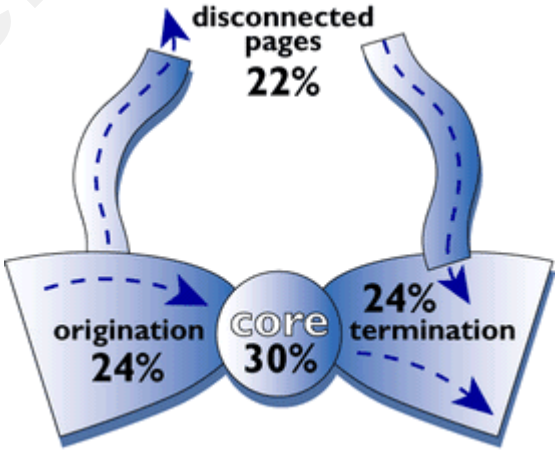


Figure 2.1.2(b)
(IBM Almaden Research, 2000)

The central piece of the structure composing of all web pages that can reach one another along directed hyperlinks forms the heart of the web and is commonly called “strongly connected component” (SCC). The SCC appears as the “knot” of the bow tie and contains about one third of all web sites. This category of web pages usually is pages that web surfers can easily travel between each other via hyperlinks.

The second piece of the web image is the upstream, also known as “origination”, consists of web pages that can reach the SCC, but cannot be reached from it; web pages that belong to this origination category possibly are new web sites that have been added to the web. Those new pages might have inserted some relevant hyperlinks to existing famous web pages but not yet discovered and linked by others. This constitutes approximately one quarter of the total web pages.

"Termination" or the downstream pages can be accessed from the SCC, but do not link back to it. Almost another one quarter of the web sites is categorized under this region, for example corporate web sites that only contain internal links.

Tendrils are “disconnected” pages that can only be connected to “origination” but are not accessible to or from the connected SCC. This portion of web structure constitutes almost one fifth of the whole web.

2.1.4 Competition for Hyperlinks on the Web

Discussion so far has highlighted that the existing web structure is formed without central planning, as a result of a bottom-up distributed process (Pennock et. al., 2002).

Eventually, based on the current distribution of hyperlinks structure, the web has exhibited its self-organization capability. This distribution is attributed to a process called "preferential attachment" or simply “rich-get-richer” (Kleinberg, 1998).

However, recent work by David Pennock et al. (2002) has proposed a simple generative model, which argues that new and poorly connected web sites can compete in the “winners take all” phenomenon. The model incorporates component that capture the two common behaviours of web page authors or webmasters, which indeed providing uniformity to the overall web environment.

In normal circumstances, web page authors insert hyperlinks to pages that they are aware of because those sites are popular. Undoubtedly, this first habit has led to the rich grow richer situation. At the same time, authors sometimes also add links of pages, which they are personally interested or relevant not just because those sites are famous, largely independent of popularity. This is the second identical behaviour of webmasters. By following this natural habit, most of the webmasters have unnoticeably taken a step that balances the “rich-get-richer” process.

In relation to Malaysia web pages, some local web page developers creating new pages might have also inserted links and so automatically attach their pages to particular portion of web structure. In such a situation, the pages to be found might spread outside the Malaysia “community” pages. This introduces great challenge for the traversing tool during its discovery.

2.1.5 Other Characteristics of the Web

In addition to the distribution of the number of links (to and from) a web page that obeys a power law over many orders of magnitude (Kumar et. al., 1998, Barabasi and Albert, 1999, Broder et. al., 2000, Albert et. al, 1999), other aspects of the web characteristics including distribution by domain (Woodruff et. al, 1996), users navigation behaviours (Catledge and Pitkow, 1994) and the depth to which a web user surfs (Huberman et. al., 1997) etc have also been studied. The findings display significant degree of regularity in the web structure (Pennock et. al., 2002) and directly relate to web traversing process.

2.1.5(a) User Navigation Behaviours

It is not difficult to reckon that Internet users usually perform two common online activities: browsing and searching. Bates (1989) in his work on "berry picking", differentiates the two activities: browsing is normally perceived as open-ended, less specific and no particular goal in mind; in contrast, searching can be loosely defined as closed task with specific answers. He then further depicts that browsing and searching are not mutually exclusive user events; a user's searching process is a constantly evolving process through browsing. More importantly, the various ways used to accomplish these Internet tasks sometimes produce interesting patterns with hidden information about the web that adds to the understanding of its structure.

According to a three-week survey (Catledge and Pitkow, 1994) done by Georgia Institute of Technology, hyperlinks were by far the preferred method of traversal, accounting for 52% of all document requests; second, accounting for about 41%, was

the “Back” command; following in order of popularity were “Open URL,” “Favorites,” “Forward,” “Open Local,” “Home Document,” and “Window History”.

The survey indicates that users typically did not know the location of documents prior to searching. Most of the time, users relied on other heuristics to navigate and reach to a specific document. Furthermore, most users did not select items in the “Favorites” and “Window History”. It seems that they either preferred using “Go To” or did not know how to use this interface technique.

Instead of thinking that the current way users navigate the net are determined by the web structure, it is logical to view from a different angle, where the existing web structure might indeed be greatly affected by the users’ behaviours. This is justifiable because most of the web sites are purposely designed and gear towards increasing the users’ convenience, making the browsing and searching easier and simpler (Bates, 1989). In other words, achieving greater users’ satisfaction decides how the web pages are arranged, and in some way affects how the existing web is organized.

2.1.5(b) Depth of Surfing

How deep does a user normally surf? This is an interesting question, yet difficult to answer, because the Internet users are coming from all walks of life with assorted habits, interests and intention of using the web. Fortunately with the recent years of research effort, several large empirical studies (Catledge and Pitkow, 1994, Huberman et. al, 1997) have revealed useful information on common patterns of users surfing behaviours.

Studying users surfing behaviours actually mean examining the decision making process of a user whether to proceed to the next page along the hyperlinks or stop at the current point. The regularity recognized from this surfing pattern yields important information for travelling the web structure to more accurately anticipate the users' information needs.

Huberman Bernado et al. (1997) have discovered some mathematical expressions to predict the surfing patterns through extensive empirical studies of different user communities. These algorithms are useful for quantifying the probability of the number of pages the users visit within a web site. The algorithms were tested and validated using representative sets of data samples.

This formulation is based on the concept that each web page a user visits has a value (V_L). Users will continue to click on the next page if the coming page is valuable too. Since the value of the next page is uncertain, the algorithm therefore assumes that the value of the page is stochastically related to the previous one. The sample data collected from 23,692 America Online (AOL) web users who made 3,247,054 clicks, at 29th and 30th November and at 1st, 3rd, and 5th December 1997 has shown that the number of pages a user visits within a web site is 2.98, almost three pages.

Another appealing observation regarding the depth of user surfing was obtained by Catlegde L. D. and Pitkow, J. E., (1994). The survey finds out that users usually accessed on average 10 pages per server. This indirectly indicates that users might give up their search for relevant information after two to three jumps of the initial homepage

(two/three navigations in, two/three out, performed two/three times). However, the survey also suggests that sites with too many links in one page do not guarantee more efficient information search because this increases the searching time. Furthermore, placement of numerous links in a page always clutter screen layout and makes it very difficult for users to surf through.

2.1.5(c) Distribution of Web Documents by Domain

Woodruff, Allison et. al. (1995) from the University of California have examined over 2.6 million HTML documents collected by the Inktomi crawler and it confirmed that after all markup had been extracted, the mean size was 4.4KB, the median size was 2.0KB, and the maximum size was 1.6MB. The distribution of document by domain appears in Table 2.1.1(c) on page 23. As illustrated in Table 2.1.1, 27% is education web pages, 20% is commercial web pages and government web pages are only 4%.

In comparison with the web structure defined by the Bow Tie Theory (Kleinberg, 1998), it is not uneasy to correspond most of the web documents from the "edu", "gov", "net", "mil", "org" and certain percentage of "com" with the SCC; whereas the remaining percentage of web pages from "others" and "com" matches the "origination", "termination" or "tendrils" of the web structure.

Table 2.1.1(a)
Distribution of Document by Domain
(Woodruff *et. al.*, 1996)

Domain	**# Of HTML Documents**	**% Of Total**
Other	1,064,318	41%
Com	516,709	20%
Edu	698,616	27%
Gov	117,125	4%
Net	113,595	4%
Mil	14,734	1%
Org	89,939	3%
Total	**2,615,036**	**100%**

''other'' includes all domains other than the given top-level domains. For example, ''other''contains all non-US top-level domains (such as Malaysia's .my).

2.1.6 Some Benefits

Apart from providing more effective web traversal, knowledge of web structure and its characteristics are initiated and driven by some research areas below:

i. Increase the effectiveness of e-commerce (Almaden Research, 2000)

Through the design of more effective browsing, advertising, measuring and modeling, e-commerce sites may decide to use different strategies for attracting online customers from various regions. For example, an "origination" web site will have to increase its efforts to be easily found by search engines. Once the site is linked to the SCC, its strategy may then shift to other traffic-generating measures.

ii. Provide guidelines for web interface designs and usability (Catlegde and Pitkow, 1994)

Several characteristics of the web suggest some useful guidelines for designing the web page with greater usability and direct the attention of web audience to “must see” information.

iii. Analyzing the behaviour of web algorithms that make use of link information (Butafogo and Scheneiderman, 1991, Mendelzon and Wood, 1995, Carriere and Kazman, 1997, Kleinberg, 1998, Brin and Page, 1998)

Avoid search engines that use link information in ranking algorithms from suffering of link "spamming" intended to create an artificial increase in number of visitor hits.

iv. Predicting the web structures evolution (Kumar et. al., 1998)

With these findings, researchers can now develop new models to study the growth of the web and possibly predict the emergence of new and yet unexplored phenomenon on the web. While some pages may evolve into the SCC, new pages will continue to be created in all three other regions.

v. Advancing in web mining algorithms

To achieve more complete coverage and able to develop more advanced mining approaches to capture useful information base on the characteristics of web structure such as distribution by domain (Woodruff et. al, 1996),

users navigation behaviours (Catledge and Pitkow, 1994) and the depth to which a web user surfs (Huberman et. al., 1997) etc

vi. Improve information retrieval (IR) on the web

Combine the understanding of web structure with some IR active research areas like web page content analysis, result sets ranking algorithms, and evaluation methodologies to achieve better web retrieval results (Kleinberg and Lawrence, 2001).

These advantages when related to the analysis results in chapter 5.0 may imply some possible suggestions to the development of local web pages.

2.2 Traversing the Web

In the previous section (section 2.1), several striking outcomes (Kumar et. al., 1998, Barabasi and Albert, 1999, Broder et. al., 2000, Albert et. al, 1999) in the research of web structure have been reviewed. Familiarity with the web structure is useful but not sufficient for developing an automated traversing tool of the web. In a traversing tool, the most integral component would be the traversing engine. Therefore, this section examines some existing traversing approaches used in the recent web searching technology. In addition, differences between web search and conventional search are compared and analyzed in a few aspects.

As briefly covered in the introduction (chapter 1.0), many existing search engines have their proprietary traversing systems at the back end (Suvellan, 2003(a)). The main duty of such systems is to gather raw information from the web to populate their databases.

The information found is then stored and indexed before accessed by users through the search function. The front-end interface that people usually enter the search queries when they visit any search engines, actually make use of information collected by the traversing tools.

Nowadays, these kinds of traversing systems have started to become familiar to people especially when the approaches are adopted by newly emerged and rapidly-growing technology. These traversing systems are used by information software agents at the present moment of time (Klusch, 1999). Besides, the information overload problem, which is becoming extremely critical, also motivates and drives the traversing tools from the backend of search engines to serve the users directly (Kowalski and Maybury, 2000). Similarly, in the process of collecting Malaysia web pages, an automated tool is believed to be of great help.

2.2.1 Web Search

Regardless of manual or automated, eventually what do searching Malaysia web pages mean? In general, searching can be loosely defined as a process for mapping users' specified needs with the information available (Kowalski and Maybury, 2000). More specifically, the web search or Internet search, as described by Barfourosh et. al. (2003), consists of two typical methods that people regularly employed: (1) clicking and following hyperlinks through browser and (2) query through the search engines in the form of keywords. These two methods contextually match with the two most common Internet users' activities termed "browsing" and "searching" denoted by Bates (1989).

Web search in this dissertation refers to both contexts; It is used interchangeably and each applies respectively to the traversing process and evaluation in later stage. Due to the broad definition of Malaysia web pages used here, traversing process of collecting Malaysia web pages is more likely to carry the meaning of browsing, which is normally perceived as open-ended, less precise and without specific objectives. On the other hand, "searching" context, which is a closed task where specific answers are needed to evaluate the effectiveness of traversing approaches in precision of pages found.

Besides looking from the Internet activities or tasks aspect, closer examination into ordinary searching (in the context of information sciences) clears up vagueness of web search in the same way. Although web search has similar objectives with ordinary searching but each has a different mechanism and structure to represent information.

The information structure of the Internet is called hypertext and differs very much from traditional information storage data structures in format and use. The outstanding feature of hypertext structure allows one item to reference to another via an imbedded pointer; each separate item is called a node and the reference pointer is called a link; and each node is displayed by a viewer that is defined for the file type associated with the node (Kowalski and Maybury, 2000). In fact, the concept of hypertext has been around for more than 50 years (Kowalski and Maybury, 2000) but it was the web that opens up better opportunity and makes more efficient use of this concept. As Mark Lager (1996) depicts, "The web uses hypertext, a protocol or common language, to jump easily between files; the web opens the publishing arena to anyone with a computer. These

hypertext links join databases, files, sounds and pictures; texts, library catalogues, songs, video, and more are now available to the computer-literate."

Unlike the orderly world of conventional library collection, information on the web is chaotic, often not organized and jumbled up without clear boundary or separation. Mark Nelson (1965) names this scenario as information anxiety - the overwhelming feeling one gets from having too much information or being unable to find or interpret data (Nelson, 1965). Even though there are increasing numbers of topic-oriented portals or online directories exist in the recent years, yet it is still a long way to achieve an optimum solution for organizing such a huge collection. As Brian Pinkerton (1994) states, "the World Wide Web is decentralized, dynamic and diverse; navigation is difficult and finding information can be a challenge" (Pinkerton, 1994).

Hypertext structure is not the only apparent difference. More significantly, the legacy knowledge representation method that assumes information takes document-like embodiments (Svenonius, 2000) varies from the web, both conceptually and technically. Legacy methods assume discrete documents that persist through time; whereas, web documents are often products of dynamic scripts, database manipulations and caching or distributed processing; (Terrence, 2000). Furthermore, the collection of the web is not static but highly dynamic, with new documents being added, deleted, changed and moved around from time to time.

Apart from the two discrepancies above, the rate of growth and size of the web also prohibits the classic ways of searching information like what commonly is used in the

library. Published information in libraries is stored and organized on microfilm rolls and catalogue cards, on magnetic tapes or disks and usually can be located by means of label (Hartner, 1981). But, these are inefficient and even inapplicable in the Internet age.

The web, a fast-growing information source, with its contemporary structure and representation, has encouraged the current tendency of automated web search tools to replace labor-intensive methods such as manual search using catalogue. In fact, there is no apparent difference shown between searching Malaysia web pages and any ordinary searching on the web. The underlying concepts and physical activities performed are more or less identical. Therefore, development of an automated tool can be a straightforward implementation of existing approaches.

2.2.2 Traversing Approaches

The design and development of major components in chapter 3.0 generally characterize the usual traversing practice. Typical process of traversing the web often starts with predefined web addresses and downloads the web pages accordingly; then, for each page, it extracts its URLs in order to be followed later in a specific manner (Barfourosh et. al., 2002), as illustrated in Figure 2.2.1 (i) on page 31. The specific manners here are usually referred to various traversing approaches.

Among all the steps shown in Figure 2.2.1(i) (page 31), select/implement-traversing approaches have the most vital impact to the overall traversing process. Ultimately, the purpose of implementing some traversing approaches is to determine the URLs that should be followed rather than blindly going around the web to search for information

of interest, in this case Malaysia web pages. The types of information intended to search always affect the choice of traversing approaches. In general, two types of information can be gathered along the traversing process: definite information (to meet particular information needs) and common information (to proliferate search engines database) (Barfourosh et. al., 2002), and this depends on whether a browsing or searching the process is intended.

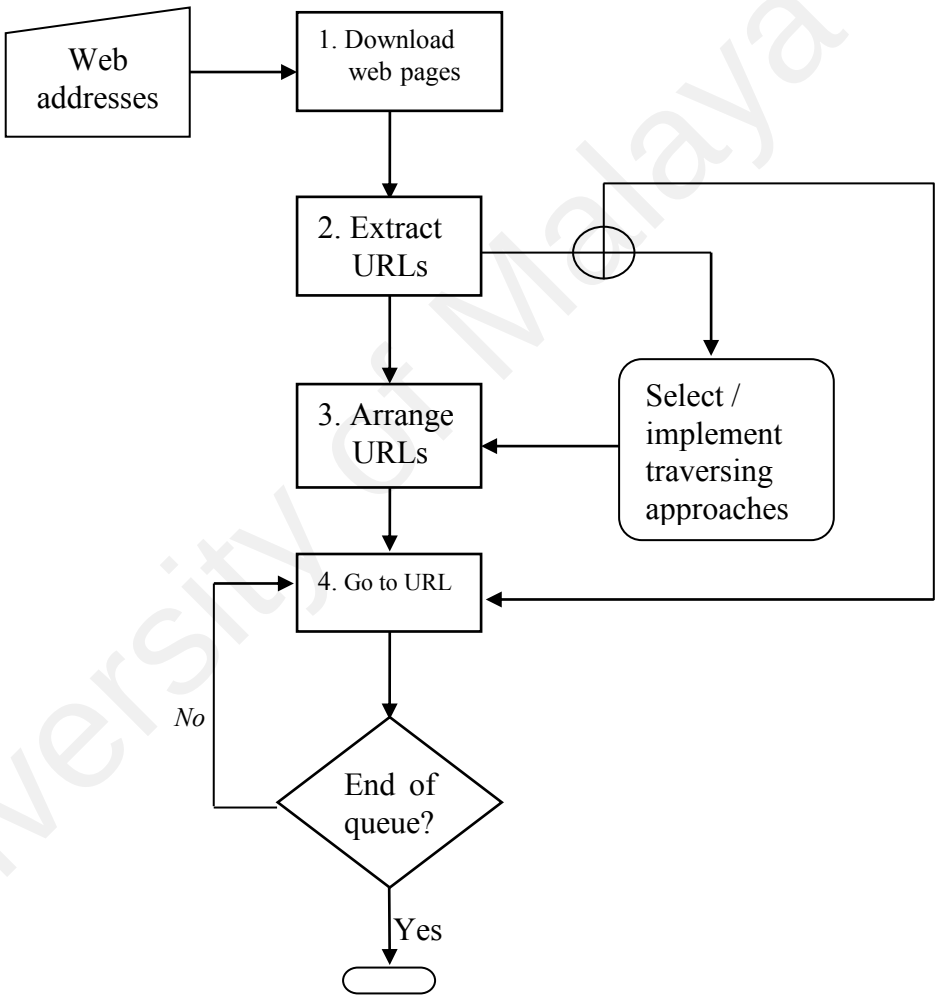


Figure 2.2.1 (a)
Flow Chart of Traversing Process

Besides the types of information, there are other factors that can influence the selection and implementation of traversing approaches. Since 1993, traversing approaches has

been undergone several progression driven by the rapid growth of Net users (MIT, 2003), which demand more sophisticated way to address new information needs. Consequently, wide range of advances has been proposed for enhancement. Until today, those approaches are broadly divided based on two major aspects: link structure and web content (Barfourosh et. al., 2002), as depicted in Figure 2.2.1 (ii).

Approaches, which use the link structure as guidance to find the path through the web, analyze relation between web links (as described in section 2.1); whereas, approaches that use web content to serve the same purpose, analyze text within each page (Barfourosh et. al., 2002).

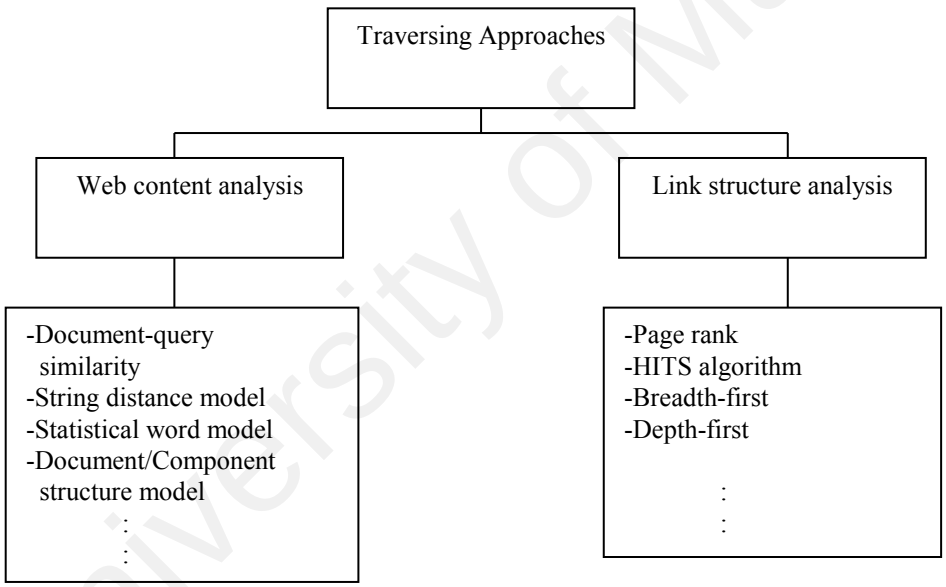


Figure 2.2.1 (b)
Classification of Traversing Approaches

2.2.2(a) Web Content Analysis Approaches

There are several algorithms that fall under the two categories of traversing approaches (as per Figure 2.2.1 (ii)). Content analysis approaches analyze and process the content of web pages downloaded from the initial input (predefined web addresses). Document-

query similarity, string distance model, statistical word model, document/component structure model are several more frequently used measurements to determine the importance of web pages. The URLs extracted from the downloaded web pages are then rearranged according to the weighting (ratio of importance) that is derived from the algorithms. In-depth explanation for the inner working of such web content analysis algorithms is beyond the scope of this dissertation because hyperlink structure is the central point of the work.

2.2.2 (b) Link Structure Analysis

The core concept of link structure analysis is to find the importance of web pages or determine their relatedness to a particular topic (Barfourosh, 2002) by using the hidden information indicated by the hypertext links. The results are then used to anticipate the traversal route.

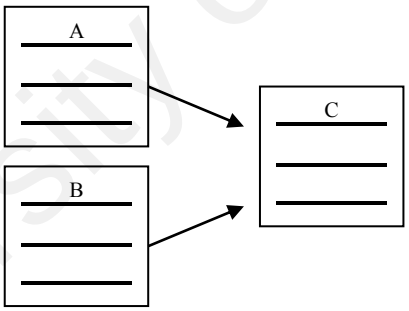


Figure 2.2.1 (b)
Example of Outlinks and Inlinks for a Web Page

When analyzing link structures in any web pages, the numbers of "incoming links" and "out-going links" are two significant features that can be observed. For example, as illustrated in Figure 2.2.1 (b), assuming web page A and B include the web address of web page C in their page respectively; hence, web page C is said to have two in-coming links, one from web page A and another from web page B. At the same time, these in-

coming links of web page C is actually one of the out-going links for web page A and B. Page and Brin (1998) point out that, " every web page has some number of out-links and in-links. We can never know whether we have found all the in-links of a particular page, but if we have downloaded it, we know its entire out-links".

Since it is almost not viable to have a collection of all the in-links that point a web page, the prototype developed in chapter 3.0 only take into consideration the out-links.

Breadth-first, depth-first, Page Rank and HITS algorithm are among the link structure approaches that can be found under the Internet traversing topics or other closely related subject areas. Although the work of this dissertation is mainly based on breadth-first and depth-first, synopsis of the Page Rank and HITS are also covered in this section to appreciate their contributions and uniqueness in the field of traversing approaches.

Breadth-first search is an earlier and general-purpose approach that many search engines employ (Barfourosh et. al., 2002). This approach does not target for finding specific information; maximum coverage is its final goal. This approach has been long defined by Cheong (1996): "breadth-first search strategy fetches as many as possible web pages to create a broad index and ambitiously aims to ensure that every server with useful content has at least several pages represented in the search engines databases".

Cheong (1996) further describes the process as depicted in Figure 2.2.1(c). The process starts with a known set of documents, examines the outbound links from them, follows one of the links that lead to a new document (discovery of new documents through

learning their identities in the form of URLs (Uniform Resource Locators)), and then repeats the whole process. He sees the advantage of this approach in not overloading web servers with "rapid-fire" requests but traverse in a friendlier manner to include as many servers as possible.

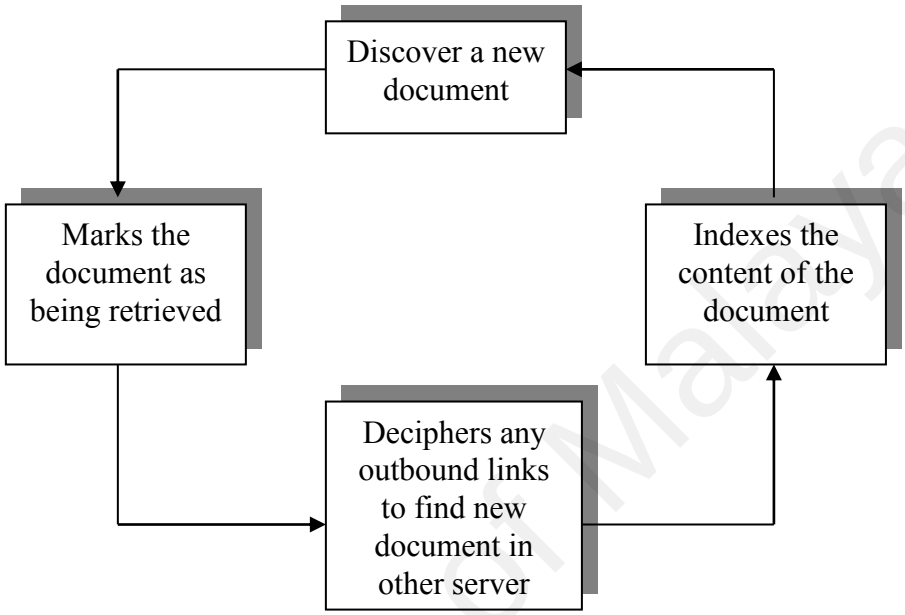


Figure 2.2.1(c)
Sequence of Actions for Breadth-First Traversal

Opposite to the breadth-first search, is a domain-specific approach known as depth-first search. Depth-first search assumes that relevant documents to a topic should be near each other in link structure. Some agent –based systems with desirable goals, implement this approach to locate relevant information in highly distributed and decentralized databases such as the web (Klusch, 1999).

According to Pinkerton (1994), the intuition behind the depth-first algorithm is that following links from documents that are similar to what the user wants is more likely to

lead to relevant documents than following any link from any document. Figure 2.2.1 (d) gives a picture of how the algorithm works as described by Cheong (1996).

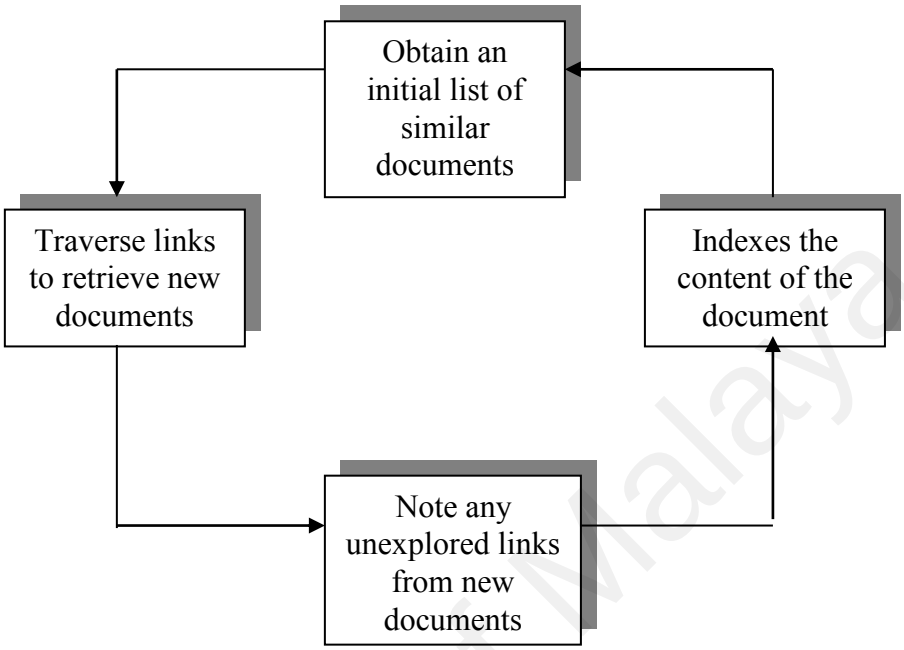


Figure 2.2.1(d)
Sequence of Actions for Depth-First Traversal

Even though major part of the development of automated traversing tool in the later stage is directly pertaining to these standard approaches but implementation aspect uses a different way to more naturally capture and represent the ordinary way people navigate the web when finding Malaysia web pages. Further description of on these has been covered in the development of traversing tool in chapter 3.0 later, where all steps involved in breadth-first and depth-first are clearly described. Table 2.2.1(b)(i) provides the summary of major differences between breadth-first and depth-first approaches.

Despite the depth-first search, Page Rank and HITS are also another two approaches more towards searching specific information. These approaches reorder the links in the

URLs queue extracted as to their predicted likelihood to lead to pages that are relevant to particular topic (Chakrabarti et. al, 1999).

Page Rank makes use of the link structure on the web and the concept is originated from the fact of how the Nobel Prizes are assigned. This approach considers highly linked web pages are more "important" than pages with fewer in-coming links (Brin et. al, 1998); same with people having more references (citation) than others who are usually awarded with Nobel Prizes. As Brin and Page (1998) explains,

> *" It is somehow different and is more sophisticated than simply counting the number of in-links of a web page. The reason is that there are many cases where simple citation counting does not correspond to our common sense notion of importance. For example, if a web page has a link from the Yahoo! Homepage, it may be just one link, but it is very important one. This page should be ranked higher than other pages with more links but only from obscure places. Page Rank is an attempt to see how good an approximation to "importance" can be obtained from just the link structure of the web". (Brin and Page, 1998)*

Very much similar to the Page Rank is the HITS algorithm, which has been briefly introduced in section 2.1.2, with respect to the self-organization characteristics of web structure. In this approach, two kinds of pages are identified from web page links: authority pages and hub pages. As per Kleinberg and Lawrence (1999), "Hubs and authorities exhibit what could be called a mutually reinforcing relationship: a good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs". On the other hand, Bharat and Henzinger (1998) pointed out a few weaknesses of HITS to represent the importance of web pages in all situations. For example, web documents generated by tools (e.g. web authoring tools, database conversion tools etc.) often have links inserted automatically. Hence the authority score of web pages inserted shoot up, but this does not mean it is more important than other

pages. In view of the limitations, researchers from the IBM Almaden centre proposed several improvements to this approach (Barfourosh, 2002, Charkarbrakti, 1999), which have led to the development trend of "focused" or "topic-specific "approaches. In terms of practicality, Page Rank and HITS have achieved commercial usage, which is encouraging, besides their successful experimental outcomes in research laboratories. Google schedules and operates their traversing schemes using the Page Rank approach and HITS is implemented by Clever search engine from IBM. Though both are working on the link structure concept, there are still some differences spotted on closer examination. Table 2.2.1(b)(ii) gives the comparison between Page Rank and HITS algorithms done by Barfourosh et. al. (2002).

Table 2.2.1(b)(i)
Comparison of Breadth-First and Depth-First

Breadth-First	**Depth-First**
General purpose searching techniques	Domain specific searching techniques
Does not target for any specific web pages	Usually locate for specific web pages
Maximum searching coverage is the goal	Accuracy of searching is the goal

Table 2.2.1(b)(ii)
Comparison of HITS and Page Rank

HITS	**Page Rank**
Can distinguish between pages with high number of in-links but not related to topic and related to given query	Blindly calculates the importance of a page according to its in-links and out-links regardless of given query
Suitable for topic driven page importance measuring	Suitable for measuring overall ranking of sites and pages and their importance from the perspective of people citation regardless of topic, estimating the popular or highly cited sites
Refined HITS considers the web page content in addition to link structure	Uses just link structure of the web

2.2.3 Some Challenges

There are at least three obvious challenges that inhibit automated web traversal to cover every page on the web, as pointed out by Barfourosh et. al (2002). Finding Malaysia pages faced the same problems too.

One obstacle is that it cannot reach all of the web documents from a single point in its graph because there is not a path from any given page to every other page on the web. Second, new pages added daily with more speed than the traversing tools capable to gather web pages. The last barrier is that some pages will be updated long before the traversal tool visits them again. These types of challenges have drawn increasing interest and attention of researchers (Broder, 2003).

2.3 Evaluation Methods

Precision and recall are two traditional effectiveness measures for information systems (Dominich, 2000). An early use of precision and recall to measure the effectiveness of an information system is traceable to the work by Cleverdon (1963 and 1964); Detailed consideration of the two concepts has been made by Salton (1968), who used them extensively in discussion of results obtained with the SMART system, a famous system in the early stage of information retrieval research. Ideally, any information systems should supply users with information, which include none that are not of interest and should not omit any that are relevant (Heaps, 1978).

In this section, the evaluation method: recall and precision are reviewed. The objective is to identity the appropriate use of these two attributes to measure the results generated

by traversing approaches in chapter 4.0, as well as to support the chapter 5.0 discussions.

2.3.1 Definition of Precision and Recall

Assume that a typical information system provides its user with M number of items in response to a search query. After manually checking all M items returned by the information system, the user decides that the total number of relevant items to him/her is P. The system is then said to have precision of P/M (Heaps, 1978). For example, a user enters a set of keywords, "world wide web", to find some information using the UM Pico search engine (http://web.um.edu.my/um/Search.htm). Say the engine turns up with 20 results and the user finds that 10 of them are what he/she wants, so system is said to have achieved 10/20=0.5 precision in this case.

Recall evaluates an information system from a different aspect than precision. Suppose after examining all the items in the database systems, the user identifies total of R items, which are of his/her interest, the information system that produces M items above is said to have recall of P/R, in respect to the search query (Heaps, 1978). Consider the same example given, assuming that the user is able to observe the entire database of UM Pico search engine and confirms there are a total of 200 items relevant to "world wide web". The engine is therefore said to have obtained 10/200=0.05 recall.

The conflict between precision and recall is not a rare issue to many information systems. Buckland and Geythere (1994) describe the different perspectives of these two concepts in which, precision represents the purity of retrieval, but recall manifest the

completeness of retrieval. They further elaborate, “Empirical studies of retrieval performance have shown a tendency for precision to decline as recall increases. Analysis of the relationships between recall, the number of items retrieved, and precision shows that there is a definable region for all feasible retrieval results. For all cases of consistently better-than-random retrieval, recall curves tend to follow an increasing curve rising form the origin, and a trade-off between precision and recall is inherent, and this is not just an inconvenient empirical finding. More generally, a trade-off between precision and recall is entailed unless, as the total number of documents retrieved increases, retrieval performance is equal to or better than overall retrieval performance thus far. There is a fundamental relationship between precision and recall which, for a given model of recall, constrains the behaviour of precision."

2.3.2 Measuring Traversing Effectiveness Based on Precision and Recall

Using precision and recall concepts as effectiveness indicators of traversing approaches is not an innovation. As Cheong (1995) emphasizes on issues faced by web indexing agents, "a good indexing scheme aims for high recall and precision. A large proportion of the useful documents should be retrieved, and at the same time a large proportion of the extraneous documents should be rejected". However, no details on how to measure the traversing approaches are found during the preparation of this dissertation.

Conventional precision and recall is not directly applicable in this work. There are two basic difficulties encountered. Firstly, one of the parameters, P, would require human perceptions to determine the relevancy. But it is infeasible to manually scan through the large datasets. Secondly, for the case of recall, it is impossible to obtain the value R,

which represents the total relevant pages in the entire database because in this case the database refers to the whole web. Hence, minor modification is adopted as discussed further in chapter 5.0.

2.4 Summary of Related Work and Current Work of Dissertation

The work of dissertation involves the understanding, which spans broadly into several subject areas. In view of this, the literature review was prepared with three objectives as basic guidelines in mind: (1) covers the fundamental knowledge necessary to develop a traversing tool, (2) familiarize some background of the related work, which is to be implemented in finding Malaysia web pages, and (3) preliminary evaluation of the tool developed and methods used.

Determining the scope of subject areas and limiting the depth of each selected area requires some considerations. Therefore, section 2.1 covers some recent works in exploring the web; section 2.2 reveals some existing approaches available for the development of traversing tools and 2.3 studies some common evaluation methods. Although all three sections show some independencies but their intimate relationship is apparent; in combination, it provides an overview and essential understanding for the work in later chapters.

The three sections in the literature review also represents the 3 major stages of work At stage 1, although the development of the prototype is mainly built on top of the existing approaches, but using different implementation of concept. Instead of server based, page-based is used for traversing the web in both breadth-first and depth-first (depicted

in more detailed in chapter 3.0). Stage 2 proposed a distinct perspective of collecting Malaysia web pages by implementing the recent web structure research. As many attempts to address the problem of finding web pages using multi-lingual approaches through word meaning (specially the delineation and discrimination of word sense), using the web structure context can be another interesting way. Finally, at stage 3, as a proof of concept, the classical information systems evaluation methods are modified to compare the percentage of relevancy and coverage for breadth-first and depth-first.

3.0 DEVELOPMENT OF AWTT

This section is to operationalize the traversing concepts of breadth-first and depth-first into Automated Web Traversing Tool (AWTT) development framework and consequently applied to the construction of AWTT. The use of these traversing approaches aims to take advantage of the existing structural topology of the web, which is the hypertext links structure (as depicted in chapter 2.0). Apart from that, the AWTT is implemented to investigate the automated process of browsing and searching for Malaysia web pages as well as to test the development feasibility and quality of web pages found (discussed further in chapter 4.0 later).

AWTT is more a backend engine than a front-end system that involves many users interaction. However, in order to support minimum user interaction and to make the system more presentable, a windows "look and feel" GUI is designed to replace the black and white DOS screens. Figure 3.1 (a) shows the use case diagram that illustrates 4 types of communications between AWTT and users. Ultimately, AWTT is a fully automated system. There are only two basic steps that require user's actions, i.e. initiate traversals and setting maximum traversal levels. AWTT responds by returning the traversal and evaluation results.

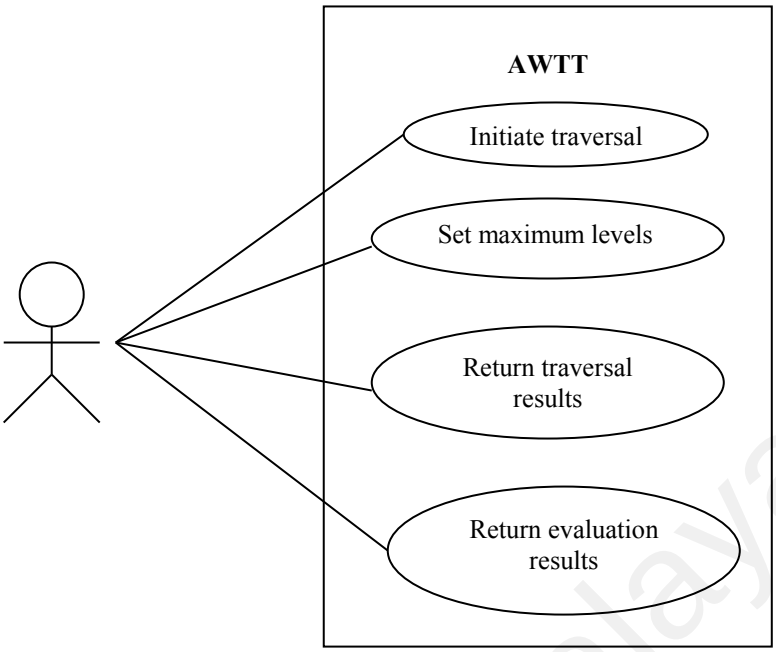


Figure 3.1(a)
Use Case Diagram of AWTT

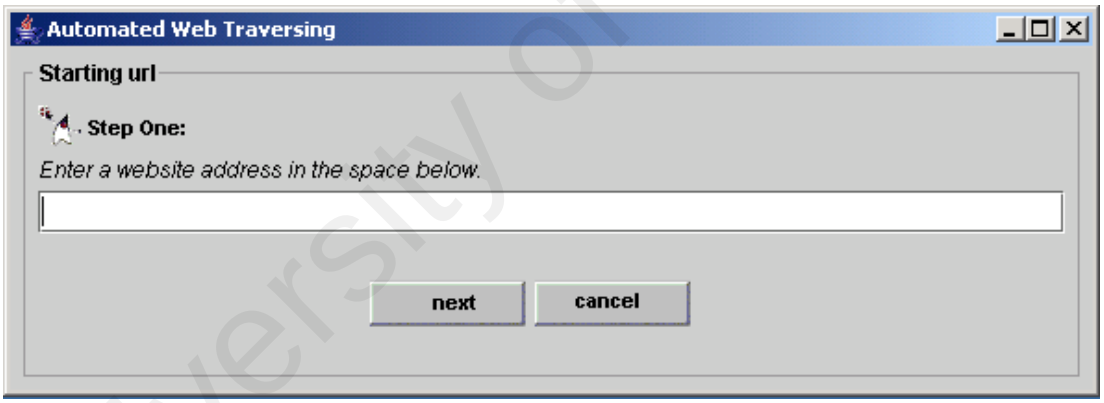


Figure 3.1(b)
Screen Snapshot of AWTT

The first step for users to start the system is by entering an URL into the start URL text field, as shown in Figure 3.1(b). Then, select the desired maximum level for traversal through the GUI in Figure 3.1(c).

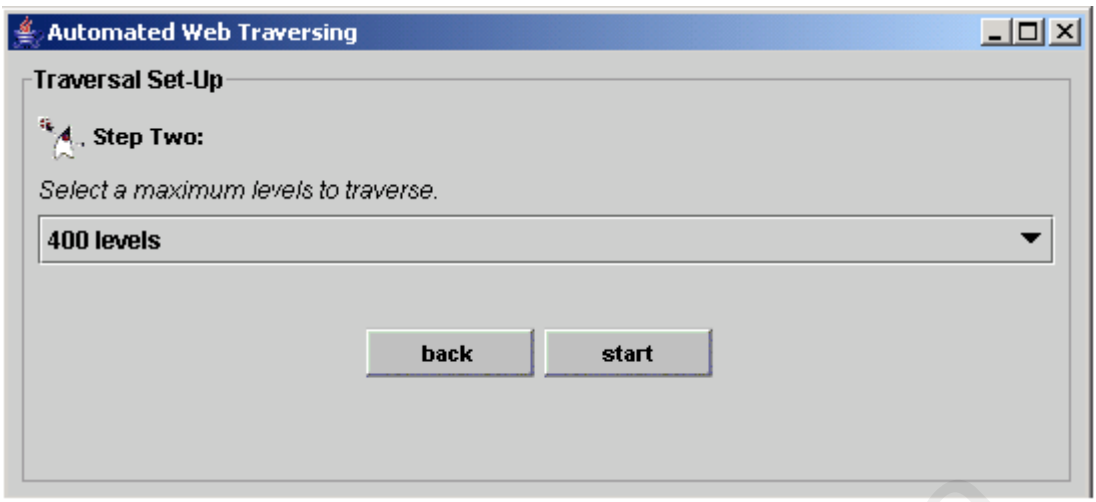


Figure 3.1(c)
Screen Snapshot of AWTT

Clicking on the "start" button launches the traversing. AWTT then starts its journey on the web until the maximum level. The progress of links fetching is displayed on the screen and all the links gathered can be found in the database. Finally, the evaluation results will also be presented to users (as per Figure 3.1(d)).

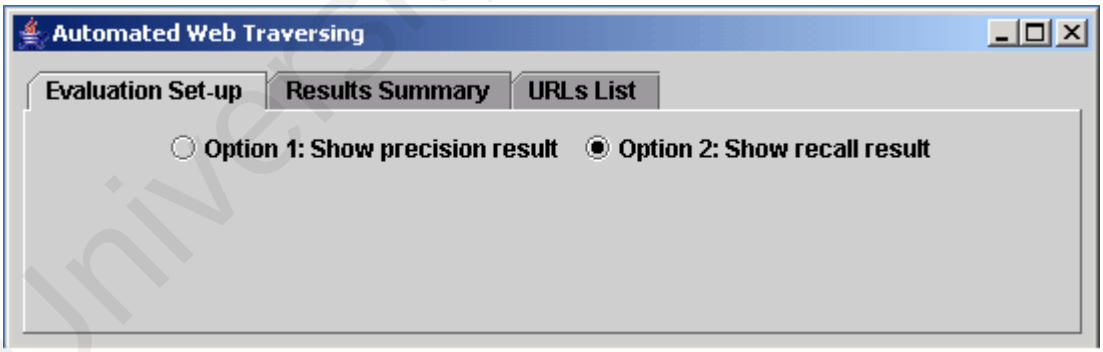


Figure 3.1(d)
Screen Snapshot of AWTT

There are several obstacles encountered during the development of AWTT, especially in automating the process. One of these problems is the presence of many uncommon protocols such as "mailto:" etc that will terminate the process. All exceptions need to be

handled gracefully to allow the automated process to continue. Besides, in extracting links from a web page, there are a lot of non-identical cases but are different syntactically. The process to deal with each circumstance is tedious and time consuming. For example, “http://www.jaring.my”and “http://www.jaring.my/” are the same URL although the latter has an extra “/” character. In scheduling the next URLs to be visited, major difficulties lies in ensuring the traversal does not exceed the maximum level and also does not repeat the visited links.

3.1 Architecture Overview

Figure 3.2(a) illustrates the simplified version for the architecture of AWTT. This architectural diagram shows an overview of major components at high level. Both breadth-first and depth-first approaches have same set of basic components to perform the functions but with different implementation (in depth description in section 3.2, page 56). The Unified Modeling Language (UML) notations (Bahrami, A., 2000) are used to describe the functional features of AWTT.

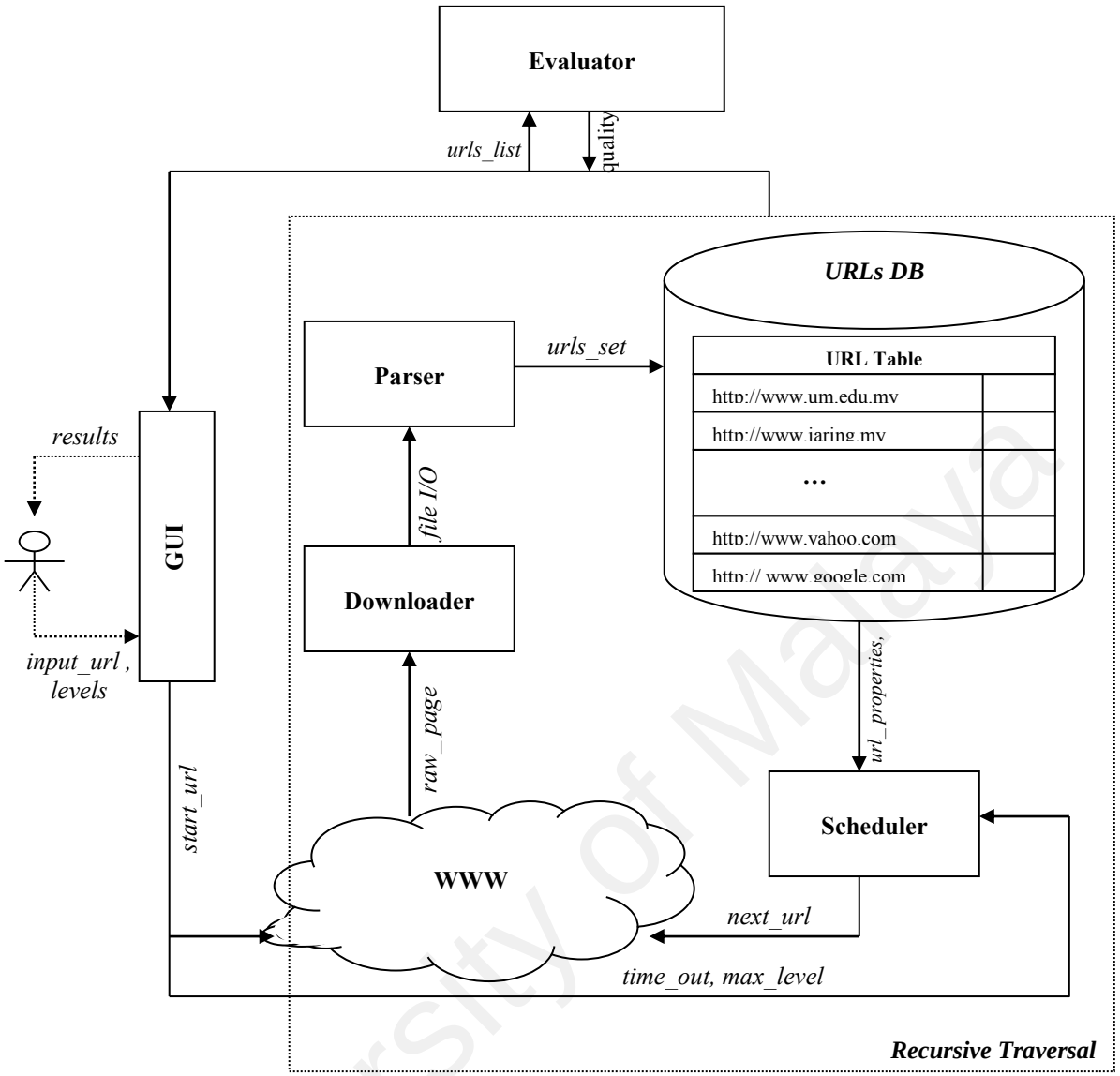


Figure 3.2(a)
Software Architecture of AWTT

The AWTT is designed following the Object-oriented (OO) paradigm and implemented in Java version 1.4. As can be seen in Figure 3.2(b), AWTT consists of 4 main classes: GUI (Graphical User Interface), Downloader, Parser, and Scheduler. All these classes have a one to one composition relationship with AWTT; in addition, each class maintains an n-ary relationship among themselves as shown in Figure 3.2(c). Besides, there is one more class, Evaluator, having the attribute of *urls_list,* associates with all

the classes above (as per Figure 3.2(c)). The naming convention for those components is also specially selected to be in line with the basic concept of OO, whereby the design extends to include ideas and terms closer to its applications. Functions of each class and its inner working are disclosed in the following subsections.

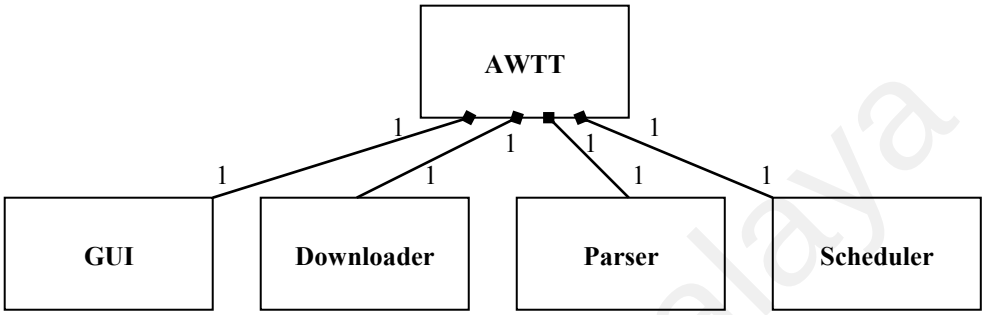


Figure 3.2 (b) Composition Relationship of AWTT‘s Classes

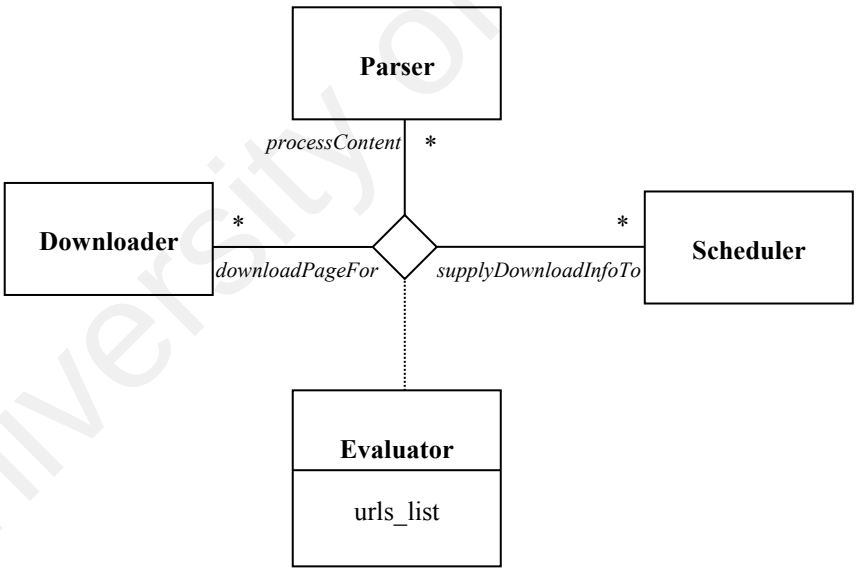


Figure 3.2 (c) n-ary Relationship of AWTT‘s Classes

3.1.1 GUI (Graphical User Interface)

Figure 3.2.1 shows the GUI class diagram. It provides common elements of GUI for the traversing process. Through this GUI, users interact with AWTT to initiate the

automated searching and browsing. The GUI presents users with a text entry box and a drop-down menu as per Figure 3.1(b) and (c).

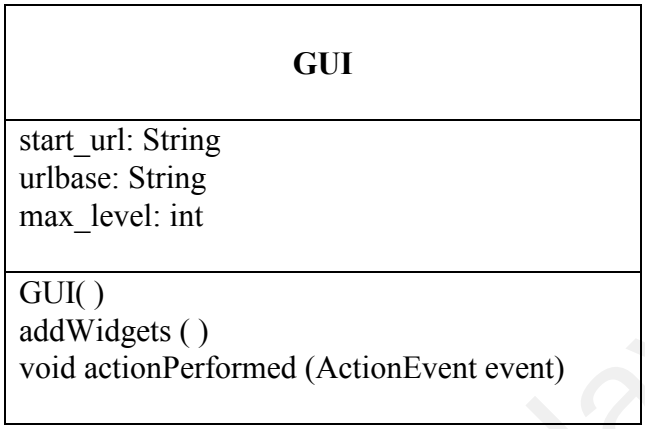


Figure 3.2.1
GUI Class Diagram

The drop-down menu contains a list of options including traversing approaches, the maximum levels to be traversed. Clicking on START will begin the automated traversal process by calling the Downloader class. However, the process would also be terminated abnormally by clicking the STOP button.

After completing the traversing task according to users input attributes, the GUI processes the results and support users as an output channel to view the URLs fetched back by AWTT. As described, Figure 3.1(a) illustrates the relationship among Internet users and use cases within AWTT.

3.1.2 Downloader

The role of this Downloader includes establishing the Internet connection for AWTT, downloads the web page as the URL indicated and automatically converts the

downloaded page into standard flat files for further processing, as indicated by methods in Figure 3.2.2. In addition, it also keep track of some web page properties such as host name, last update date and content type to be used in later stage. After extract links in the *start_url*, this Downloader retrieves the list of URL from the database and opens a connection to each URL. It then fetches the HTML document of the page and stores it as a flat file in a temporary directory repeatedly until the traversal process is ended or terminated.

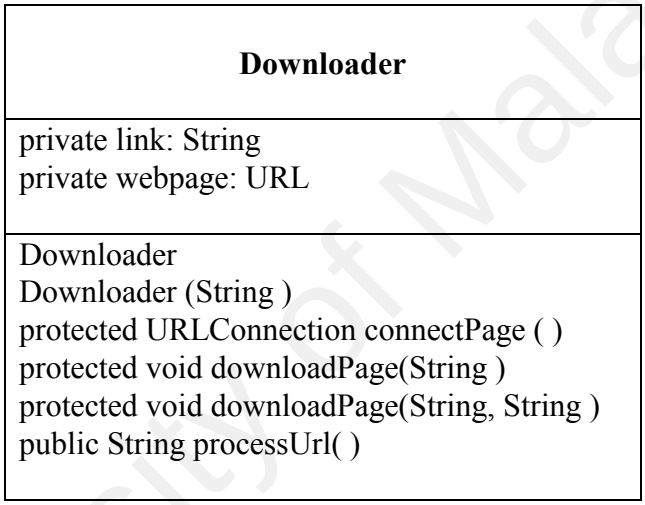


Figure 3.2.2
Downloader Class Diagram

3.1.3 Parser

This parser carries out content processing on web page in standard flat files format. It parses through the HTML documents supplied by the Downloader in the temporary directory, with the ultimate goal in search of any links within the files. The program serves for four main tasks: (1) determine the types of web page to be processed (2) extracts hypertext links from the web page (3) update the URLs database for new links to be traversed (4) update the retrieved links into the *QueueTable* in the URLs database

for next traversal. There is a slight difference in the choice of URLs to be updated into the database depending whether a depth-first or breadth-first approach is running at that time as highlighted in section 3.2. Figure 3.2.3 depicts the data and methods encapsulated with the Parser class. *(Please refer to appendix (d) on page 180 for sample raw page)*

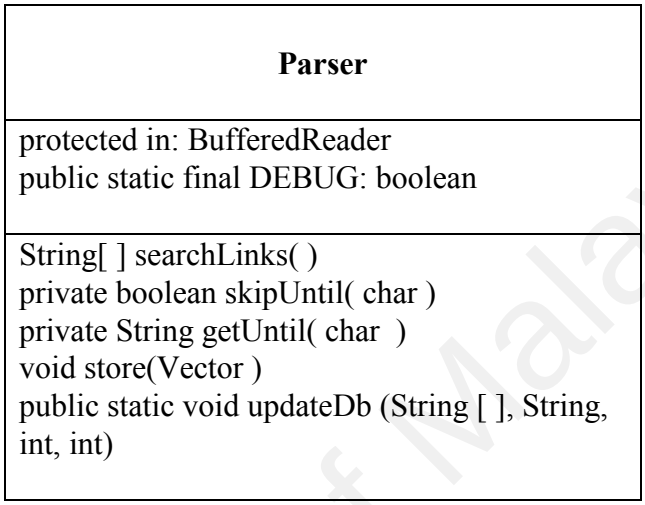


Figure 3.2.3
Parser Class Diagram

3.1.4 Scheduler

The Scheduler is a component that indirectly communicates with users by planning the traversal route based on the users settings when the process is initialized. It keeps a simple profile that register users' choices that are be transformed into parameters to monitor the overall process. One of the main methods is to retrieve the first URL stored in the *QueueTable* of the URLs database by executing a query to select the URL of the minimum ID number. The *checkLevel* is another important method of Scheduler. It is called recursively in order to check for the current levels of links that have been parsed at after each traversing cycle. This is to ensure that the AWTT will run in a controlled automated environment when user intervention is not present. The scheduler also

provides different traversal mechanisms for both Breath-first and Depth-first (see section 3.2).

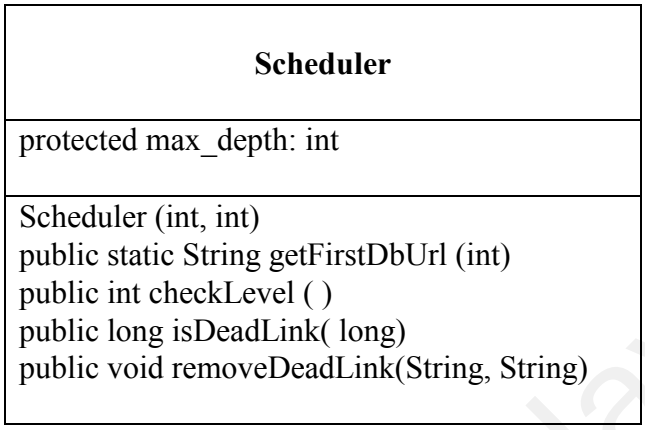


Figure 3.2.4
Scheduler Class Diagram

3.1.5 Evaluator

The Evaluator is more like an independent component. It links up with the system through supporting the evaluation for the quality of discovered web pages. Different criteria can be considered but precision/recall with respect to hypertext links are metrics used in AWTT.

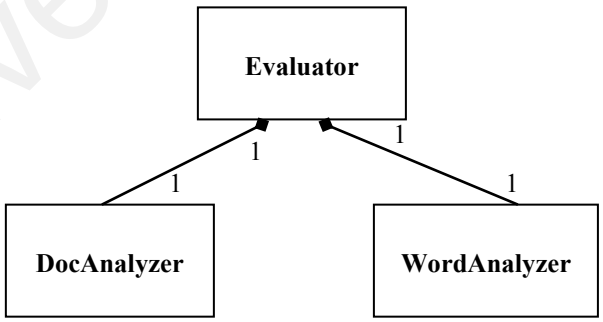


Figure 3.2.5(a)
Composition Relationship of Evaluator's SubClasses

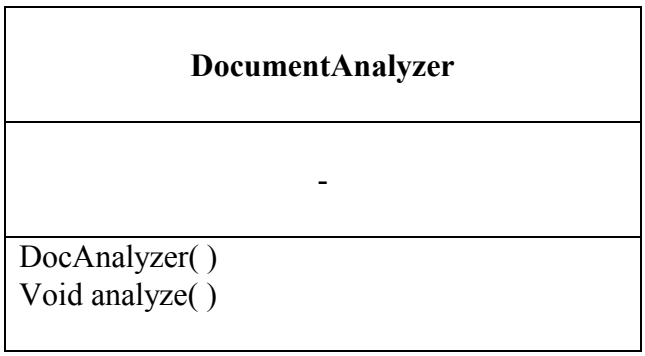


Figure 3.2.5(b)
DocAnalyzer Class Diagram

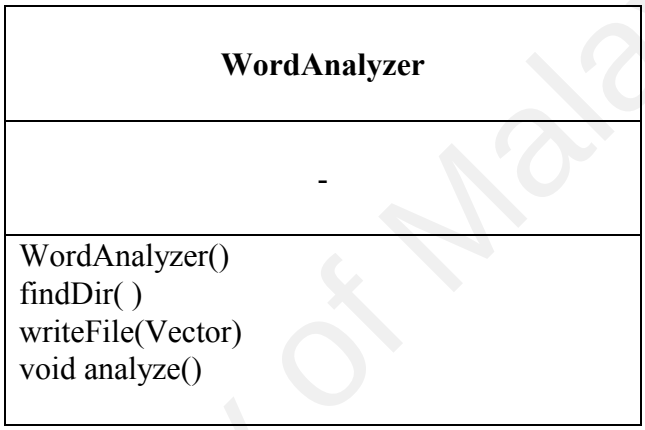


Figure 3.2.5(c)
WordAnalyzer Class Diagram

Evaluator has two sub classes as shown in Figure 3.2.5(a). The functions of these two subclasses as in Figure 3.2.5(b) and (c) are the physical implementation of calculating the recall and precision described in section 2.3.2.

3.1.6 URLs Database

The database is to store the URLs extracted and provides information for scheduler to manage traversing route. To allow AWTT to be more flexibly tested on different operational platform, for example windows 98/2000/XP/NT, Unix and Linux etc, the

database employs a standard interface to relational databases called Open Database Connectivity (ODBC) by using the JDBC (Java Database Connectivity) –ODBC Bridge to allow common database SQL operations such as add, remove, update records etc. The URLs database is a collection of three MySQL tables that is continuously processed throughout the AWTT navigation process:

(a) UrlTable
(b) QueueTable
(c) *LevelTable*
(d) *VisitedTable*

3.1.6(a) UrlTable

The *UrlTable* is an important table. It keeps track of all the extracted pages. The design in the *UrlTable* is as Table 3.1.6(a) below. The *FromID* field is designed to store the *ID* of the referrer URL. This is for the convenience purpose of the Evaluator to compute the quality of web pages found. The *FromID* has the same value with the *ID* stored in the *FromID* field of *the QueueTable*.

Table 3.1.6(a)
Design of UrlTable

Field name	***Description***	***Field Type***
ID	Indexing identification	Integer
URL	Address of web page	String
FromID	The ID of the referenced URL where the link is extracted	Integer

3.1.6(b) QueueTable

The *QueueTable* is a temporary table that stores the links extracted from the visited URL. The links are then retrieved and the Downloader goes through the links, and the

link becomes a visited URL and deleted from the *QueueTable*. The *QueueTable* update is a continuous process in a minimum amount of time. Table 3.1.6 (b) shows the fields in a *QueueTable*.

Table 3.1.6(b)
Design of QueueTable

Field name	***Description***	***Field type***
ID	Identification of web page not visited.	Integer
URL	Address of web page	String
FromID	ID of the page where the link is extracted	Integer

The *ID* field is same with that of the *UrlTable ID*. Besides, the *FromID* field is necessary to pass the referrer URL information to the *UrlTable* and also to be used in keeping track of the traversing levels.

3.1.6(c) LevelTable

The *LevelTable* (Table 3.1.6 (c) on page 55) stores information regarding the current levels that the Parser has processed. The *LevelTable* consists of the first node for every level and the level number as well as the name can be retrieved. The *LevelTable* only stores the first link in every level in order to monitor the maximum level to be traversed. Although the URL itself is not important and is a redundancy of the *UrlTable* information but it is still required to monitor for referral links.

Table 3.1.6(c)
Design of LevelTable

Field name	***Description***	***Field type***
ID	Traversing level that the AWTT has traversed.	Integer
TotalUrls	Total number of outlinks in a web page	Integer
URL	The first node at each level.	String

3.1.6(d) VisitedTable

The *VisitedTable* has only one field, the URL that the AWTT visited. Although there may be a lot of links queuing in the *QueueTable* for traversal but some of them could be unreachable. Furthermore, the *QueueTable* is only a temporary table. Hence, the *VisitedTable* is very important to record all the actual visited links.

Table 3.1.6(d)
Design of VisitedTable

Field name	***Description***	***Field type***
URL	The first node at each level.	String

3.2 Algorithm

As described above, each component is treated as an object class and is encapsulated with data and methods individually. They are instantiated and accessed through individual methods to perform the tasks and in combination accomplishing the traversal for AWTT.

The typical traversing process illustrated in 2.2.2(i) forms the basis for the logical design of AWTT. Download web pages, extract URLs and arrange URLs are the integral processes of traversing, which are equivalent to the Downloader, Parser and Scheduler of AWTT.

In terms of breadth-first and depth-first approaches, page-based implementation was used here. This is dissimilar with the ordinary server-based implementation as described in section 2.2.2. Figure 3.3(a),(b), (a)(i), (b)(i) exemplify the detail of these approaches.

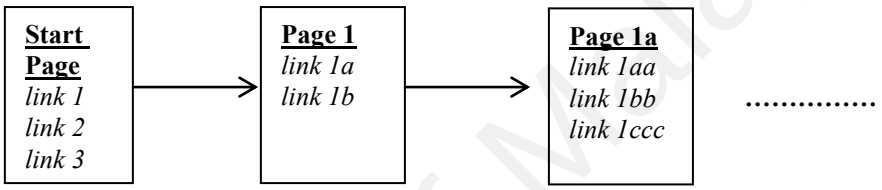


Figure 3.3(a)
Breadth-First Traversal of Page Based Implementation

From example the Figure 3.3(a), breadth-first traversal starts with *link1*, followed by *link1a* and then *link1aa,* without considering in which server those pages are located. Whereas, in a normal breadth-first web traversal, the traversal route is based in where the pages are located. For instance, if it happens that *link1a* in the *page1* located in the same server with the page containing it, then *link1b* will be traversed next if it is not in the same server. Figure 3.3(b)(i) further illustrates the sequences of this page traversal.

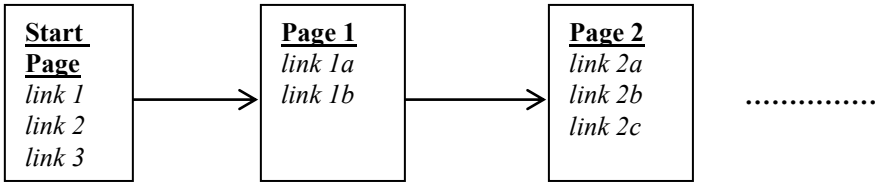


Figure 3.3(b)
Depth-First Traversal of Page Based Implementation

Similarly for depth-first traversal, no physical location of web pages in the server is considered for deciding the traversal path. From Figure 3.3(b), depth-first traversal visits all the links within a page. So, the route in the above example would be *link1*, *link2*, *link3* and so on, as shown in Figure 3.3(b)(ii).

This page-based implementation represents the human browsing and searching process, which seldom takes into account the actual location of web pages situated, links are usually followed sequentially or randomly as how they are positioned in the web page containing them.

Figure 3.3(c) simplifies the interaction of components within AWTT and the sequential flow once the traversal is initializing by users. From the sequence diagram, it is clear that there are two manual input data i.e. the *input_url* and *level*. These two inputs are used as attributes to decide the *start_url* and *isMaxLevel*. *isMaxLevel* condition is used to monitor the recursive traversing process. Whereas, *isDeadLink* condition is governed by a built-in mechanism in the Scheduler to check the validity of URLs and to determine the *next_url* to be visited.

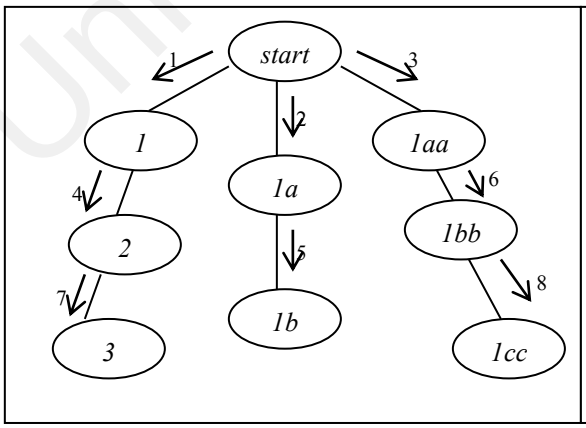


Figure 3.3(a)(i)
Tree Diagram of Breadth-First

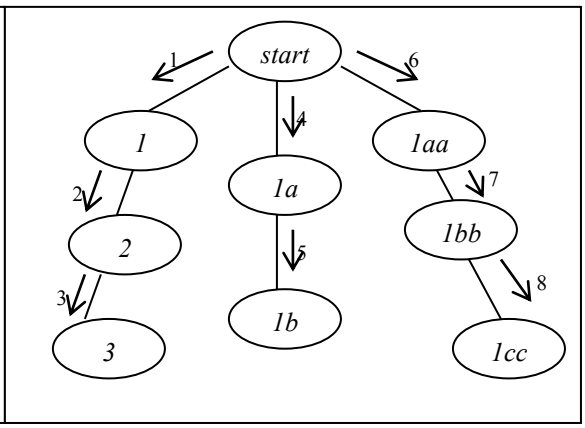


Figure 3.3(b)(i)
Tree Diagram of Depth-First

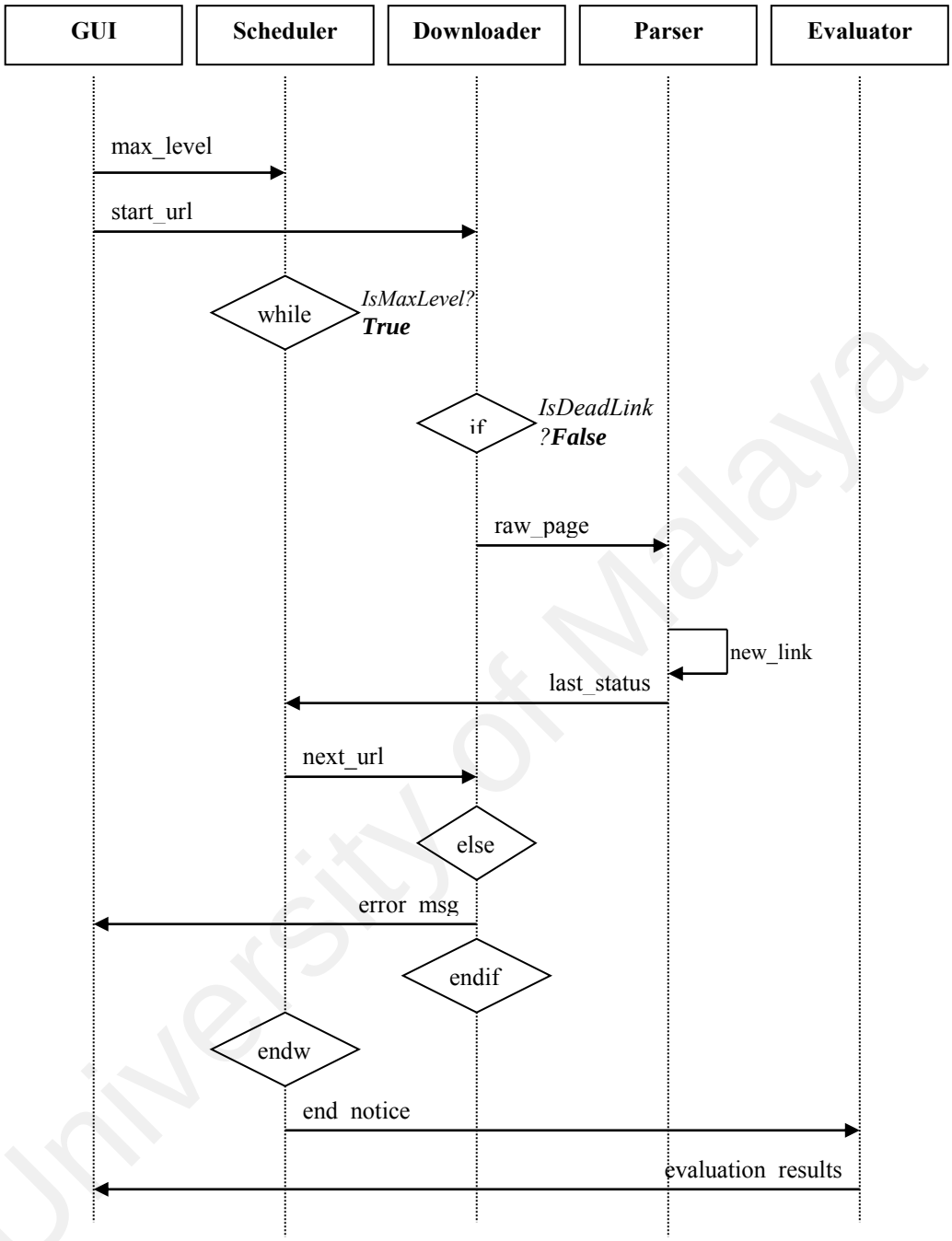


Figure 3.3 (c)
Sequence Diagram of AWTT

Besides, as depicted in section 2.2.2, the select/implement approaches have the most critical effect that can guide the traversal more systematically instead of blindly surfing

around the web. Two approaches are implemented for AWTT i.e. breadth-first and depth-first to complete the investigation objectives stated in chapter 1.0. Overall, these two approaches have the same process flow. The inner working of the AWTT is illustrated through activity, collaboration and statechart diagrams below.

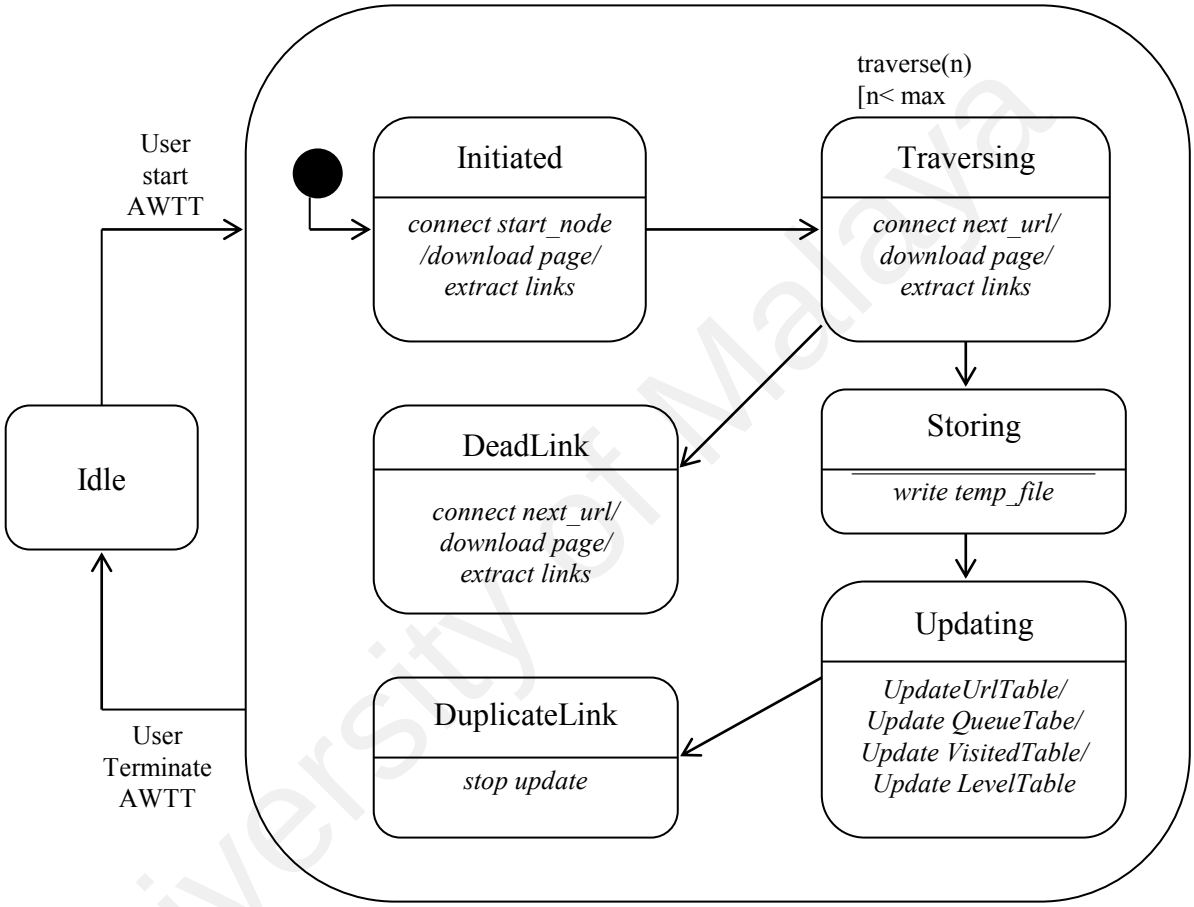


Figure 3.3 (d)
Statechart of AWTT

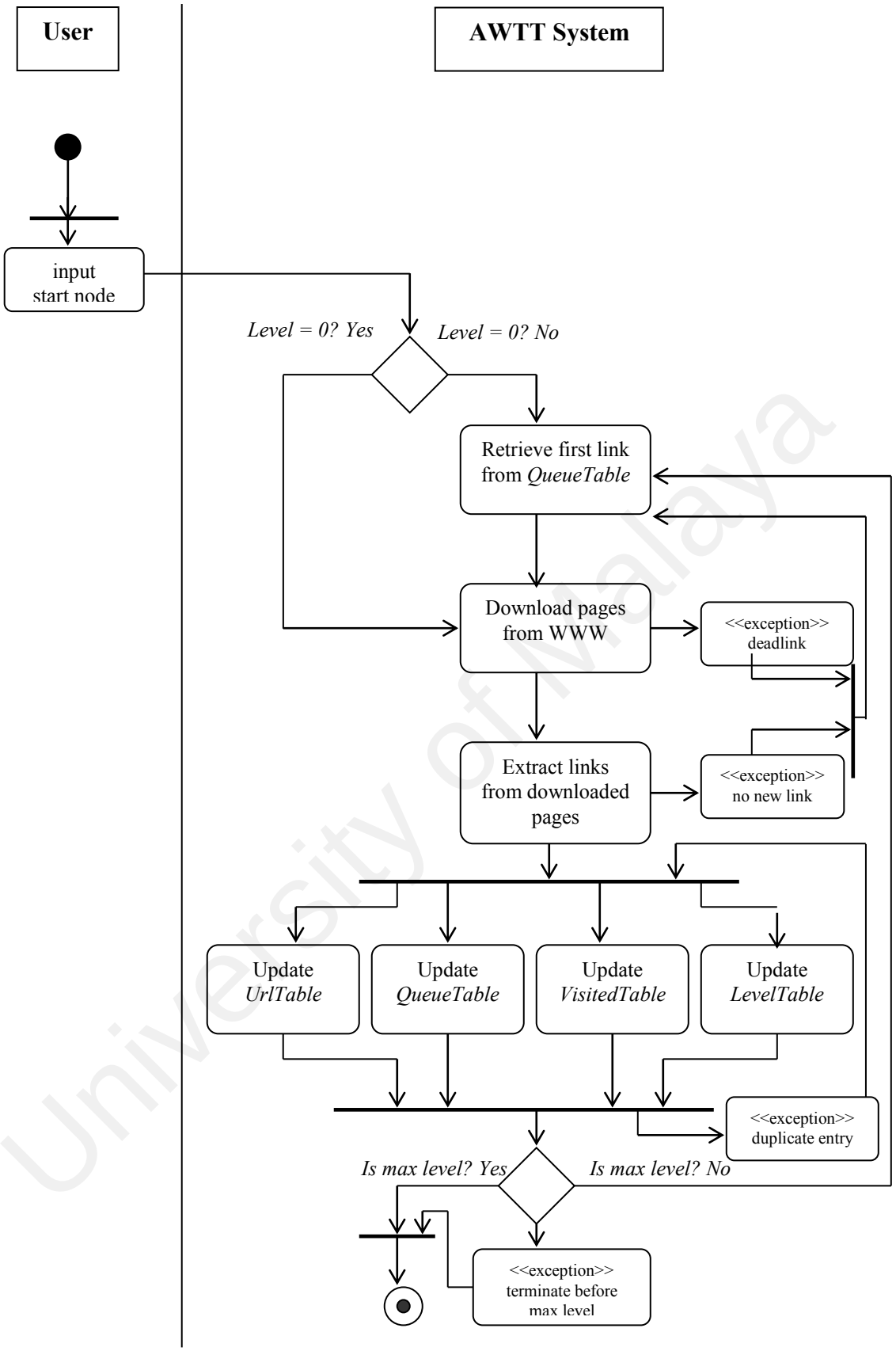


Figure 3.3 (e)
Activity Diagram of AWTT

3.3 Physical Deployment

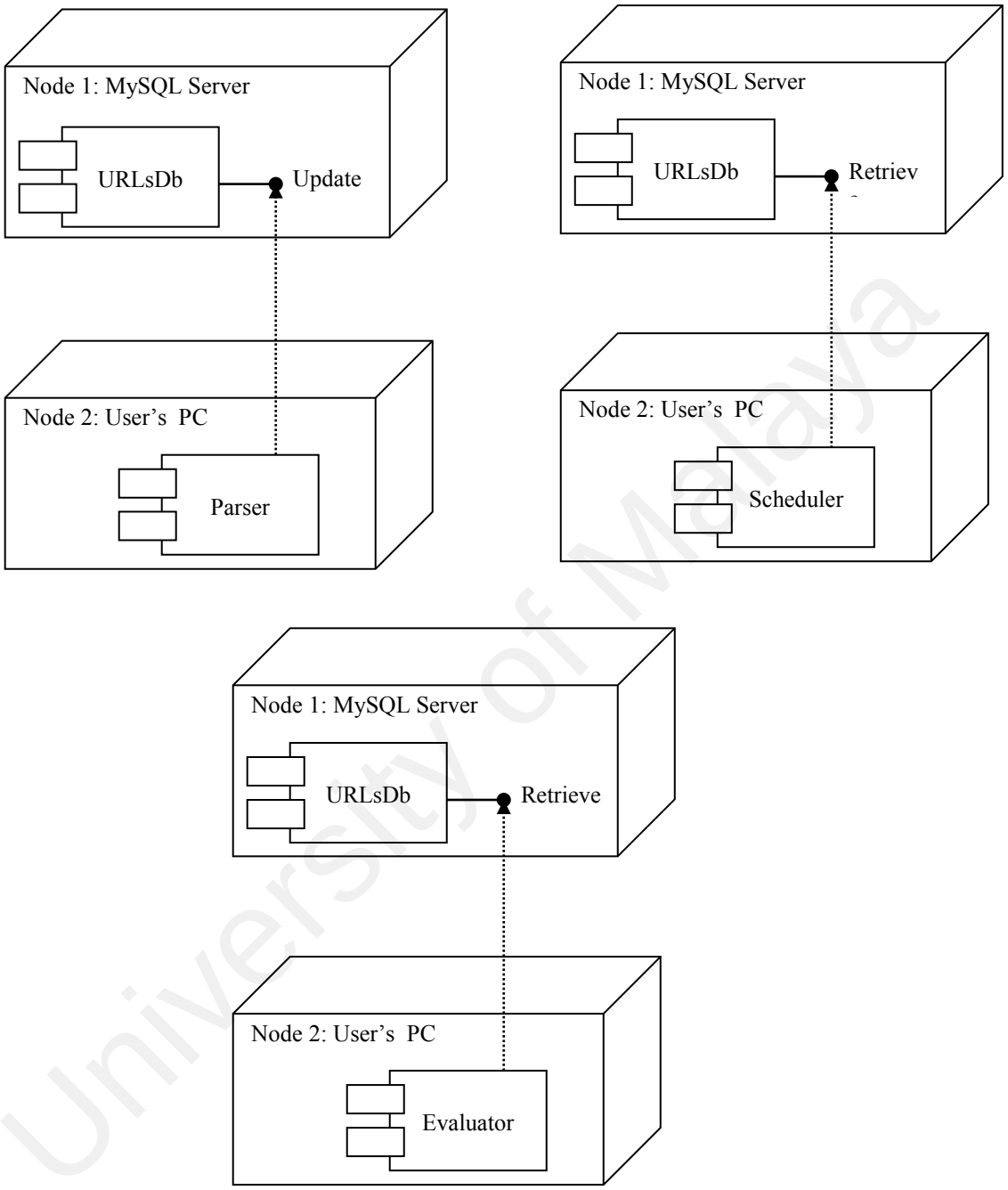


Figure 3.4
Deployment of AWTT

The physical modules of AWTT prototype are distributed on two nodes: a MySQL 4 database server and a client PC. Both are running on LINUX Redhat 8 platform, on the

same machine with Pentium II Processor, which has 256MB RAM. Figure 3.4 on page 61 shows that the MySQL server is connected to the user's PC through access of Parser, Scheduler and Evaluator to obtain necessary processing.

(*Please refer to appendix (c) on page 118 for the source codes of AWTT*).

4.0 EXPERIMENTAL SET UP

This chapter explains the experimental set-up comprehensively by elaborating on the data preparation process and some steps taken to conduct the experiment.

4.1 Objectives

As a necessary first step in any project, the main objective must be clearly recognized. Practical experiments are usually quite different than development projects, where the final goal is normally a product that meets certain specifications; most of the time, researchers conducting experiment are frequently trying to discover, establish or prove some laws of nature. (Diamonds, 1989)

In the same way, there are two ultimate goals for this experiment: (1) investigate the process of automating web traversal process and (2) compare the quality of information found using breadth-first and depth-first traversing approaches. In addition, the experiment strategy is also specifically designed to align with the research methodology and expected outcomes as stated initially in section 1.5 and 1.6.

Feasibility is another aspect to be tested apart from the two main objectives. This happens to be a hidden objective when designing the experiment because if the experiment strategy is infeasible, then it is imperative that different options need to be considered to prevent wastage of time and resources. Feasibility usually examines whether the specified properties is feasible or infeasible in the defined variable space (Diamonds, 1989). In fact, the feasibility experiment was conducted in the early stage

prior to running the actual experiment to determine 3 important properties: start node, size of data sets and maximum traversing level.

4.2 Data Collection

The data preparation process is mainly to deal with selecting suitable sample data sets. The data sets collected are only considered appropriate if they can produce results that achieve the experimental objectives stated above. The following sub-sections depict the consideration involved in deciding the start node, size of datasets and maximum traversing levels as well as several important assumptions, which have significant effects to the overall experiment.

4.2.1 Start Node

Despite a user input to manually indicate a start node, deciding a start node actually means identification of a good starting point. A good starting here can be generally seen as a node that has high probability to link with other relevant nodes in the same domain.

In order to find these good nodes, hypothesis on the web structure needs to be formed. Referring to the existing research work on hyperlinks distribution (as shown in Figure 2.1.3 (b)), the starting nodes are supposed to be one of those SCC pages, because SCC pages are web pages that can reach one another along directed hyperlinks. Only if the traversal is started with a node that is highly connected with others, then the AWTT can continuously and automatically move around on the web. Since the work aims to focus on Malaysia web pages (as per section 1.5), therefore the candidate start node must not only be able to link with other web pages but must also be relevant to Malaysia.

As described in section 2.1.3, the category of SCC pages is usually pages that web surfers can easily travel between each other via hyperlinks. Hence, one simplest way to find out those start nodes would be to make use of some existing popular search engines. By entering a search query, search engines will return many web pages as search results. However, only top ranked results are popular web pages that Internet users usually browse through. Therefore, highly ranked pages among the search results can be considered as SCC pages, as well as good starting points to launch the AWTT.

According to the SearchEngineWatch.com report by David Sullivan (2003 (b)) dated 28th October 2003, there are several popular search engines, as shown in Table 4.2.1. These search engines were evaluated based on the Hitwise rating that uses a combination of anonymous web surfing data provided by ISPs in various countries and its own panel-based measurements to determine which sites are most popular on the web. The data encompasses the surfing activities of 25 million people worldwide.

Based on the survey above, the five most popular search engines Google, MSN Search, Excite, AskJeeves and Alta Vista are selected to find the start nodes. Although Netscape and Yahoo! are also sitting at the top of the listing but they are powered by Google, and always provide exactly the same results as Google. Whereas, the rest are less suitable because some give results that are mostly commercial sites, for example iWon and MyWebSearchSearch. Furthermore, some are more document type specific such as Google Image Search (for searching image files), Yahoo!Directory (for searching personal home pages, i.e Geocities etc), Netscape White Pages (for searching white pages) etc.

Table 4.2.1
List of Most Popular Search Engines
(Sullivan, 2003(b))

Name	Domain	Share
Google	www.google.com	13.0%
Yahoo! Search	search.yahoo.com	10.1%
MSN Search	search.msn.com	7.4%
Excite	www.excite.com	1.3%
Netscape	www.netscape.com	1.2%
iWon	www.iwon.com	1.1%
Ask Jeeves	www.askjeeves.com	1.0%
Google Image Search	images.google.com	0.8%
Yahoo! Directory	dir.yahoo.com	0.7%
Netscape White Pages	wp.netscape.com	0.6%
My Way / MyWebSearch	www.mywebsearch.com	0.5%
AltaVista	www.altavista.com	0.4%
Dogpile	www.dogpile.com	0.4%
InfoSpace	www.infospace.com	0.4%
Yahoo! Yellow Pages	yp.yahoo.com	0.4%
Total		39.1%
Source: Hitwise.com for SearchEngineWatch.com		

After choosing the search engines, the keyword "*Malaysia home pages*" was entered into query box of each search engine and the URLs of the top 10 results returned are then kept as start nodes for traversing, as illustrated in Figure 4.2.1.

In Figure 4.2.1, there is one last search engine, which is not listed in Table 4.2.1 but was used to find the start nodes to begin traversing: Lycos. This is due to the rapid growth and the rate of Internet adoption of Malaysian in recent years and have definitely led to the increase of Malaysia web pages as revealed by the current ISP market and assesses future trends (Calvert, 2003). Therefore, many local web pages are still considered very new in relation to a lot of existing or well-known pages on the web. This group of web pages belongs to the "origination" category (as depicted in section 2.1.3).

Compared with all other web pages in the "origination" category, many local developed pages might have inserted some relevant hyperlinks to existing famous web pages which have not yet been discovered and linked by others. These pages are equally useful as the SCC pages to be treated as start nodes for traversal.

One possible way to obtain the URLs of these pages is to search using local search engines such as Cari.com, Skali.com, Catcha.com etc. Since the whole collection of local search engines database consist of purely Malaysia pages so local engines will not provide any meaningful results with the same query "*Malaysia home pages*" as the other five engines. In view of this, Lycos was selected because it has an international database plus a more comprehensive and local-focused collection (Hoffman, 2000), which will return both non-Malaysia as well as Malaysia web pages.

Figure 4.2.1
Finding start nodes using popular search engines

4.2.2 Size of Datasets

The target size of datasets for this experiment is about 10,000 Malaysia web pages (as stated in section 1.5). According to the latest statistics from MYNIC (The Malaysian Network Information Centre), total domain names registered under ".com.my", ".net.my", ".org.my", ".edu.my", ".gov.my" and ".mil.my" until 30th September 2003 is 43,688 (MYNIC, 2003). In this case, the figure 10, 000 is indeed sufficient to give a fair view to this experiment.

4.2.3 Maximum Level

Three types of testing to validate on the feasibility were carried out prior to the actual experiment. This is essential to determine the number of levels that could achieve approximately 10,000 web pages. After running AWTT with 3 randomly selected start nodes (*http://www.jaring.my, http://www.newmalaysia.com, http://www.um.edu.my*) for 100 levels, 500 levels and 1500 levels each node, it was found that on average each 1000 levels will generate roughly 200 pages. Therefore, it was decided that the experiment needs at least 50 start nodes with 1000 maximum levels each, hence giving a total of about 50,000 pages (50 x 200) at the end.

4.2.4 Assumptions

As per William J. Diamonds (1989): "scientists and engineers can only perceive reality, with varying degrees of clarity, by means of hypothesis, theories and experiments. Even the most precise experiments and the most closely argued theories can only be meaningful within certain tolerances, confidence levels, and assumptions." Likewise, there are a few assumptions involved in this experiment:

(1) Definition of Malaysia web pages

Malaysia web pages here refer to any existing pages on the Internet that describe Malaysia in any aspects that is related to the social, economy, politics, geography, history, business or news of the country, either in Malay language or non-Malay languages.

(2) Ranking Capability of Search Engines

Search engines have various proprietary ranking algorithms that show different accuracy and relevancy (Sullivan, 2003(c)). In this study, all engines were considered to have similar technological strength and all top ranked result sets provided are assumed to be good web pages as start nodes.

(3) Search Query for Start Nodes

The term "*Malaysia home page*" is assumed to be a generic phrase that is appropriate for finding any web pages with its content relevant to Malaysia as defined in (1).

4.3 The process and experimental platform

Figure 4.3 below summarizes the experiment process. The data preparation stage starts with collecting Malaysia web pages from six major search engine, Google, MSN Search, Excite, AskJeeves, Alta Vista and Lycos. Then the top 20 search results returned by all engines were recorded. The results returned usually contain duplicates web pages and some deadlinks. Those results were eliminated and finally 50 web pages were selected as start nodes for the subsequent stage.

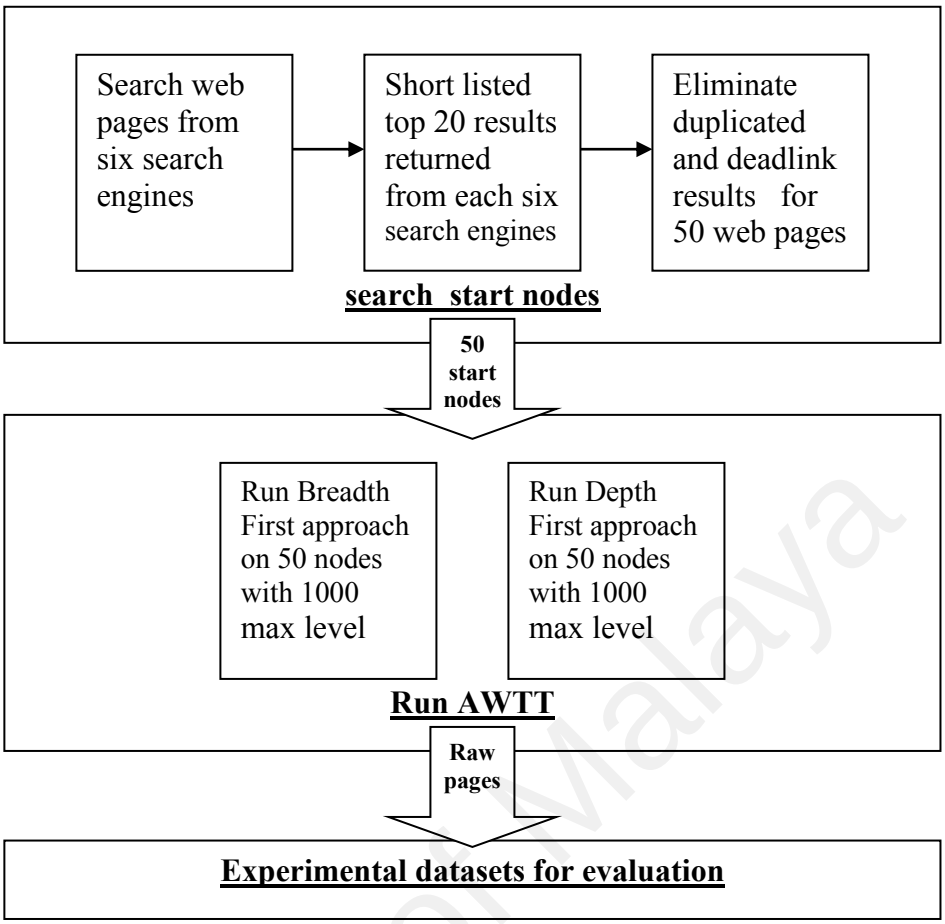


Figure 4.3
Summary of Experiment Process

The 50 start nodes were used as starting point for the AWTT to traverse the web with maximum of 1000 levels in both approaches i.e. breadth-first and depth-first approaches. The experiment was operated on the same development machine, which has a Pentium II Processor, 256MB RAM, running on LINUX RedHat 8 Operating System, MySQL 3 and with the Internet connection speed of 100mbps.

This experiment was started on 31st October 2003, 1:47p.m until 10th November 5:30p.m, with an average time of 5 hours per day. Total of 8020 raw web pages were collected using breadth-first and depth-first approaches *(please refer to*

appendix (a) on page 97 and appendix (b) on page 108 for summary of experimental reports and summary description for whole collection of datasets). These datasets were then further analyzed, compared and evaluated. The analysis outcome and thorough discussion is described in chapter 5.0.

5.0 DATA ANALYSIS AND DISCUSSION

The analysis present in this section corresponds to particular collection, which are the 50 sets of representative selection of web pages gathered in chapter 4.0. This section also contains description regarding the representative items and methodologies that are chosen to form the basis of analysis prepared for the evaluation of AWTT with two traversing approaches breadth-first and depth-first. The pre-processing to find the representative items as well as the analysis methodology has been implemented in the Evaluator (depicted in chapter 3.0).

5.1 Analysis Methodology and Preprocessing

5.1.1 Precision

Precision is one of the important aspects to be assessed in this work. As defined in section 2.3.1, precision represents the relevancy of results provided by an automated information system. In this case, precision measures the percentage of web pages that the users think is relevant to Malaysia, fetched by AWTT using breadth-first and depth-first. Although the analysis domain is focused on Malaysia web pages but obviously this analysis also responds to a general question below:

- Which approach finds pages that are more relevant?

In view of the data size of 8020 web pages, it is less practical to manually scan through each web page content to analyze. The analysis methodology used here is fundamentally built according to the manual item clustering process that is inherent in any library or filing system, where someone reads the item and determines the category

or categories to which it belongs (Kowalski, 2000). As per common understanding, ordinary human effort in grouping usually involves examining the object, recognizing distinguishing elements and finally putting similar objects together. This clustering process of web pages therefore involves implementing suitable clustering techniques to automatically group together pages that has similar characteristics, in this case elements that describe Malaysia.

Before the similar web pages can be grouped together, an important process is to identify the features of each page. The feature extraction of a web page basically involves finding the representation of the word vector or set of descriptors that best describe it. Concise representations are usually derived from the contents of more complex objects. In the case of textual objects i.e. web page, words taken directly from the page are augmented with weights and traditionally used to form a *bag-of-words* representation disregarding the linguistic context variation at the morphological, syntactical, and semantic levels of natural language (Frakes & Baeza-Yates, 1992).

Taking from the same notion above, two famous web pages: http://www.jaring.my and http://www.newmalaysia.com is selected as benchmarking Malaysia web page because they were the only two web pages, which were returned as top ranked results by the major search engines when searching for Malaysia web pages (see Table 5.1.1(a)). Words of these two pages can be treated as features that distinguish Malaysia pages from others. This is similar to the Lycos philosophy as per Mauldin (1994), in which 100 words list from several documents can be combined to produce a list of 100 words in the set of documents. Hence, the words within these pages were extracted and treated as a set of keywords that describe Malaysia web pages. This set of keywords were then

used to match with words contains in all pages gathered by AWTT. By matching the set of keywords with words found in each page, the average percentage of relevancy of web page can be calculated as following:

$$W_k \ / W_p \times 100$$

W_k - Total number of words found in web page, which matched with the keyword sets

W_p - Total number of all words found in a web page

Table 5.1.1(a)
Ranking of Jaring and newMalaysia page by major search engines

ranking of representative web page / **search engine**	**www.jaring.my**	**www.newmalaysia.com**
Google	1	3
Altavista	2	1
Excite	4	1
AskJeeves	3	2
Lycos	3	2

An additional data "cleaning" process is required to filter out "noise" that is useless and non-representative, which consist of HTML tags, common words, and repeating non-root word that carries the same meaning. So, all html tags are first removed before words are extracted from each web page. Next, all occurrences of commonly used English words known as stopwords as listed in Table 5.1.1(b) are also eliminated (Frakes & Baeza-Yates, 1992). The remaining words will then undergo a process

known as stemming to reduce them to their root forms for example, “eating” to “eat”, “travelled” to “travel”, “places” to “place” etc. In this analysis, Porter stemming algorithm is used to stem all words extracted from web pages (Porter, 1980).

After going through all processing steps above, the two benchmarking Malaysia web page yielded 255 distinct words. These set of words were used as representative item for precision analysis.

The relevancy of each page in this analysis is given by the average percentage of words matched with this keyword set, and the maximum average relevancy among the 50 groups of web pages is used as an indicator to decide a web page as Malaysia page.

Finally, the precision is computed by dividing the total number of pages that has more than or equal to the maximum average relevancy with total number of pages collected by AWTT:

$$P_w / P_g$$

P_w - Total number of web pages that is ≥ maximum average relevancy

P_g - Total number of web pages in the group

Table 5.1.1(b)
List of Stop Words

<table>
<tr><td rowspan="2">PREPOSITIONS
about above across after against along among around at before behind below beneath beside besides between beyond by despite down during for from in into like near off of on out over through throughout till to towards under until up upon with within without</td><td rowspan="2">CONJUNCTIONS
and but or nor so for yet after before when while as since until once as whenever every time because since even though although whereas while if unless therefore consequently nonetheless nevertheless however unless moreover furthermore providing provided</td><td>QUESTION
when how where who what why whom which</td><td>ARTICLES
a an the</td></tr>
<tr><td>PRONOUNS
I you she he it we they me her him it us them mine yours hers his ours theirs my your its our their</td><td>TENSES
can am are is may must have has will shall could was were had would should cannot</td></tr>
</table>

5.1.2 Recall

The process to analyze recall is much simpler as compared to precision. It does not require any representative items but concerns only the coverage of searching. As pointed out in section 2.3.1, recall is a measure of the ability of searching to find all relevant items that are in the database. In this context, recall evaluates the capability of AWTT in getting Malaysia web pages from the extraordinary huge database, the WWW with respect to the breadth-first and depth-first approaches.

Since the web is an uncontrolled environment and it is impossible that every single Malaysia web pages on the web is known, hence, one of the simplest method to compare the recall of two traversing approach is to judge against the total number of web pages that each approach is able to gather.

5.2 Analysis Result and Discussion

5.2.1 Precision

This sub section analyzes the precision analysis results derived from section 5.1.1 above, by examining the precision from each group aggregate and then comparing the results of each group in two approaches. Results of the precision analysis were provided based on the data collected from 29th October 2003 to 10th November 2003.

In discussing the cumulative precision of web pages fetched by AWTT for both approaches, it is necessary first to examine the keyword matching rates of each group, which provide the maximum average relevancy percentage of pages that is used as a threshold to determine whether a web page is related to Malaysia. As in Figure 5.2.1(a), the highest percentage achieved is 20%. Therefore, if a web page has 20% or more average relevancy then it is considered to be a Malaysia web page. Although 20% is not an ideal percentage as according to Tsunenori (2002), performance of web retrieval of the latest TREC-9 (Text REtrieval Conference), a completely automatic system (ric9dpn) at the best, whose precision is 27%. Several areas have been identified and depicted in section 6.0 to improve this aspect of AWTT for future use.

Figure 5.2.1(a) also showed the precision analysis results. The data suggest that highly relevant Malaysia web pages gathered in this experiment consists mostly of pages collected using depth-first. Lower percentage of web pages gathered by breadth-first appears to have been related to Malaysia in comparison. This result exists supports the view that breadth-first is a general-purpose approach, which does not target for finding

specific information; where as, depth-first is a more domain specific searching approach with desirable goals (as described in Section 2.2.2(b)).

When taking a closer examination on individual web pages in each group for both breadth-first and depth-first approaches, the analysis data show changes in relevancy as AWTT traverse to the levels away from the start node, with significant trend of repeated increasing or decreasing fluctuation.

Figure 5.2.1(b) shows an example of the relevancy of pages grabbed by AWTT with "*http://www.interknowledge.com/"* as start node using breadth-first. The pages have higher relevancy before the 181st web page, with many pages scoring relevancy above 25%. But after this page, the relevancy values of pages oscillate within a relatively lower band, below 15%. Then, the relevancy percentages increase again, close to 25%. Theoretically, breadth-first traversing are supposed to produce results that give continuous decreasing trend without bouncing back as reflected in the last part of the graph. The fact that significant portions of the web can be bridged by using path going through intermediate page as identified by A.Broder et. al (2000) (section 2.1.3) explains this scenario of AWTT fetching back more relevant pages after crawling far from the start node.

Although depth-first gathered web pages that yield higher relevancy percentage, but the amount of web pages gathered in collectively was small, and this relatively small number produces trend differences which are significant, 13 out of the 50 groups indicate decreasing value in relevancy corresponding to the levels of traversing, 2 out of the 50 groups show repeated increasing or decreasing fluctuation trends as with the

breadth-first results and 35 groups shows no changes. This seems to violate the depth-first assumption that relevant documents of a topic should be near each other in link structure. A collection of data that is at least 10 times larger than the existing one is expected to produce a fairer view for the analysis in this aspect.

The data also imply the existence of self-organization characteristics of the web (as shown in Figure 5.2.1(c)). Given the likelihood of topically related pages that has more links to pages within the "community" than pages outside the "community", the data reveals remarkable clusters (in terms of relevancy percentage) as described in section 2.1.2.

Figure 5.2.1 (a)
Relevancy of Web Pages

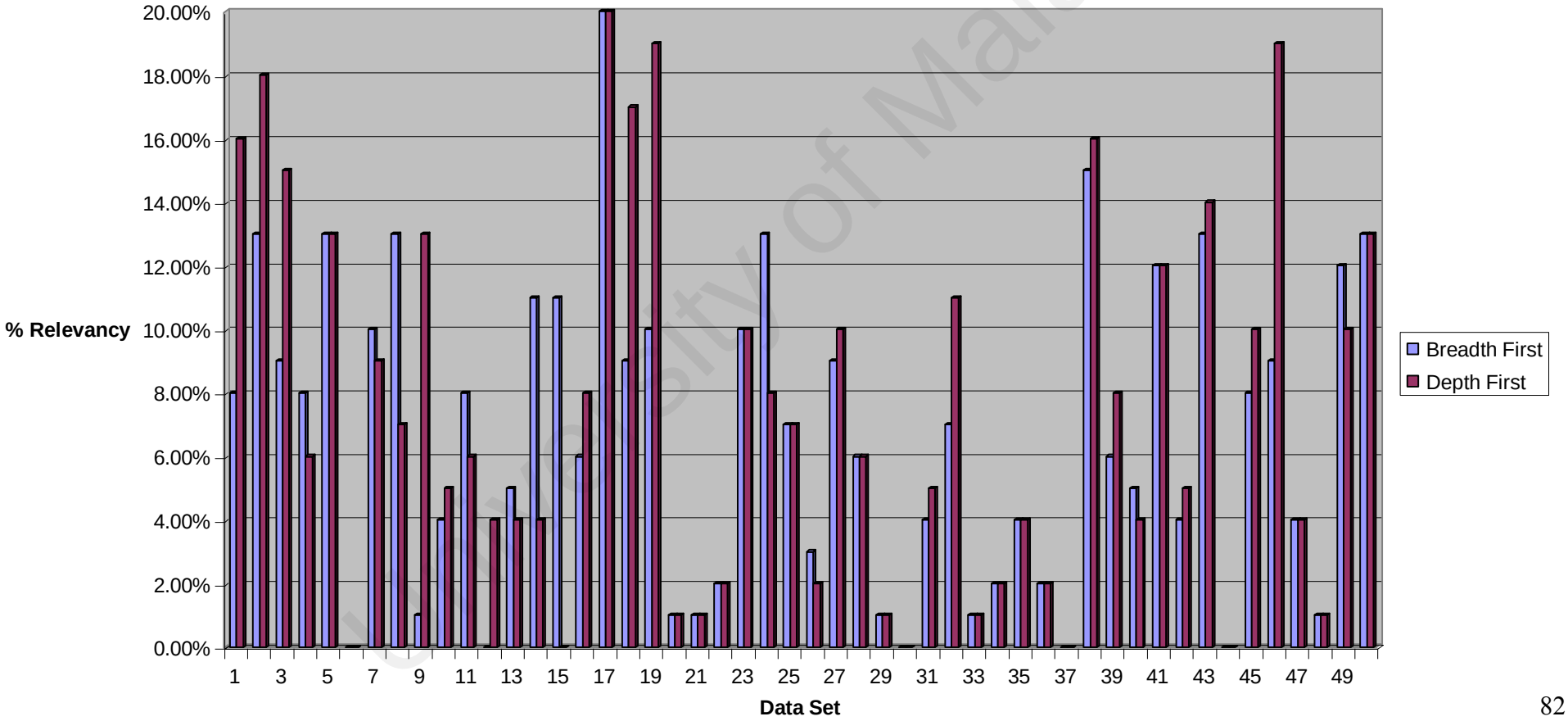


Figure 5.2.1(b)
Changes of Relevancy

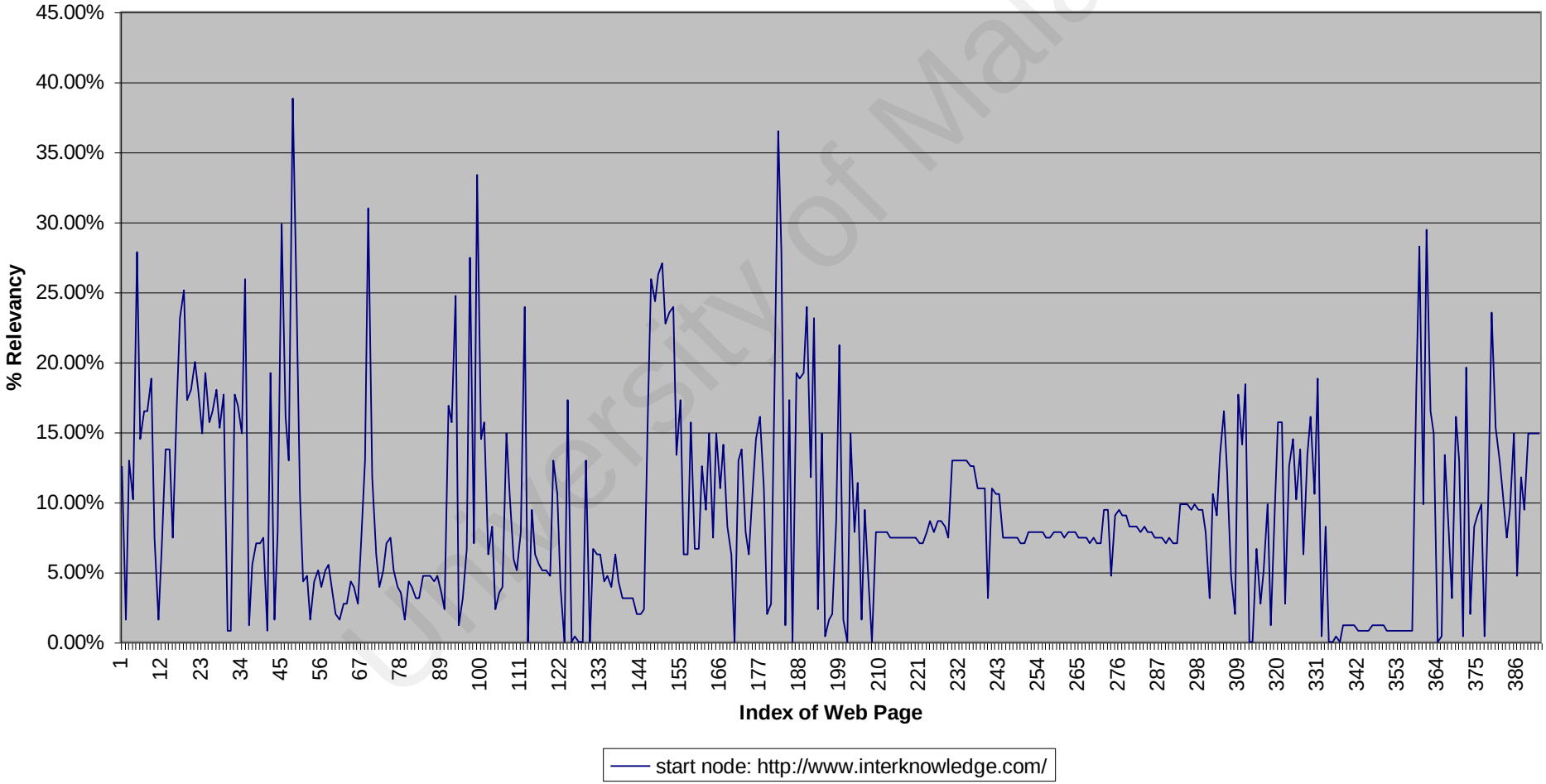
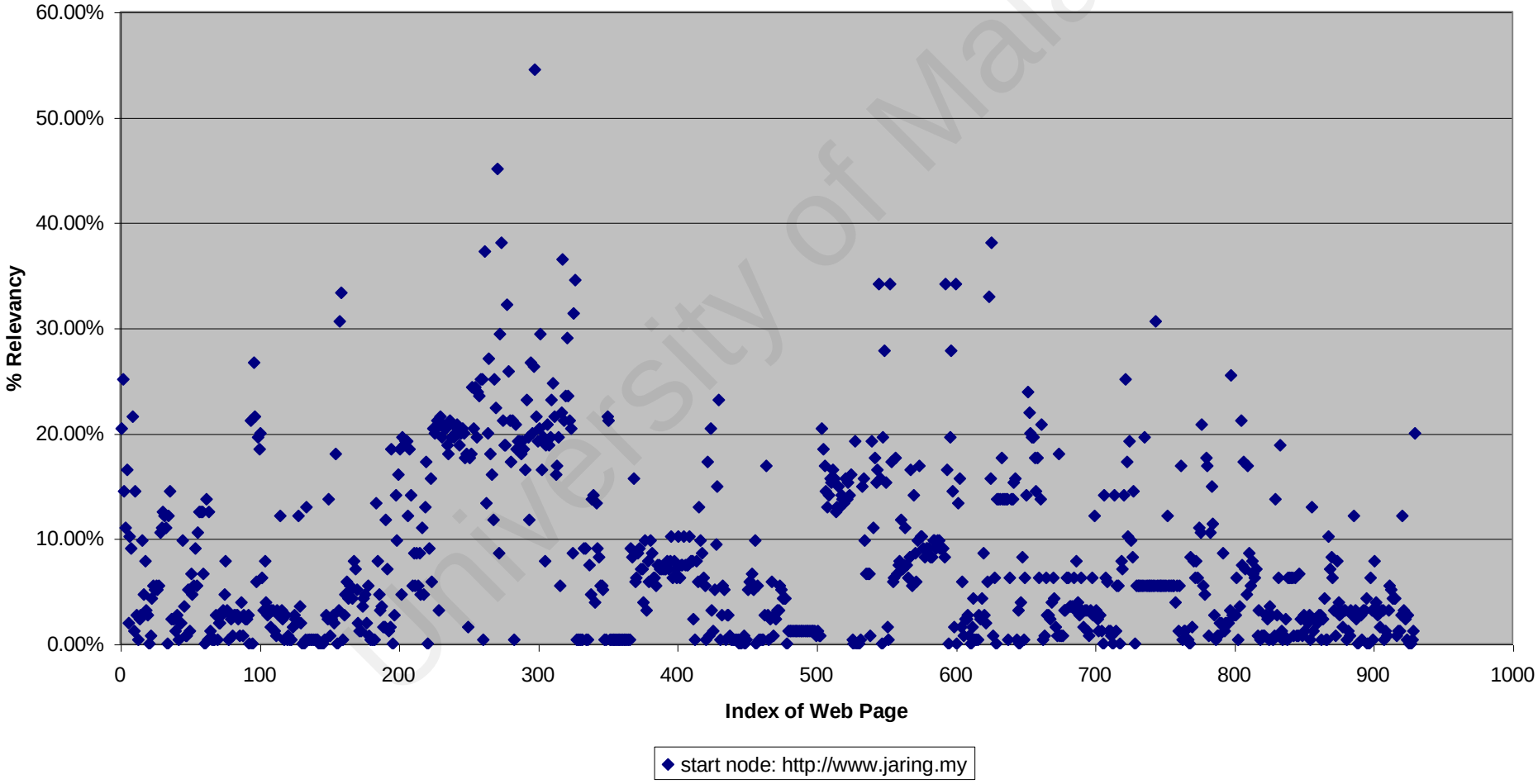


Figure 5.2.1(c)
Clustering of Web Page



5.2.2 Recall

Which approach can AWTT use to collect more web pages? This is the question to be answered by recall analysis. From Table 5.2.2, it is apparent that depth-first traversing of AWTT is far less performing than breadth-first in terms of total number of web pages collected.

This could be due to chances that are higher for depth-first approach to encounter termination pages (as describe in section 2.1.3), which are mostly corporate websites that only contains internal links. Any situation of obsolete links, corporate server downtime, heavy-loaded sites, and server time out etc can easily terminate the AWTT and discontinue the traversal.

In contrast, breadth-first approach allows AWTT to jump from one page to another without drilling down. This lead to the discovery of more new pages, which could be sitting at the origination or core portion of the web. As being identified in section 4.1, many locally developed pages might have inserted some relevant hyperlinks to existing famous web pages. Since only good sites can attract more Net users and become popular, it is logical to presume their pages have better maintained sites and experience fewer situations described above (Shi et. al, 2002). This causes the amount of pages grabbed a lot more than depth-first as Table 5.2.2 exhibits.

Another observation from the experimental data is that AWTT usually traverse not more than 4 levels. Although this result is not very encouraging but it may not be critical since reports have shown that depth of user visit within a website is 2.98, almost three pages (as per section 2.1.4(b)).

Table 5.2.2
Total Number of Web Pages Collected

Start Node	Breadth-First	Depth-First
1.	930	2
2.	248	8
3.	392	8
4.	1000	2
5.	1	1
6.	5	0
7.	211	2
8.	1	2
9.	4	2
10.	5	2
11.	110	2
12.	1	1
13.	32	2
14.	915	2
15.	960	0
16.	934	8
17.	1	1
18.	846	2
19.	20	2
20.	2	2
21.	2	2
22.	2	2
23.	3	3
24.	6	2
25.	9	5

Table 5.2.2, continued
Total Number of Web Pages Collected

Start Node	Breadth-First	Depth-First
26.	7	2
27.	1	1
28.	1	1
29.	6	2
30.	1	1
31.	1	4
32.	5	4
33.	15	1
34.	1	1
35.	1	1
36.	2	2
37.	1	1
38.	20	2
39.	4	1
40.	36	4
41.	507	7
42.	5	4
43.	517	4
44.	0	0
45.	2	1
46.	117	3
47.	1	1
48.	2	2
49.	5	3
50.	1	1

5.3 Issues Raised by the Data Sets

This section is to account for several uncertainties present in the data sets to generate a more complete view of analysis. In examining the 50 experimental data sets, it is important to examine not only the average scores for relevancy and recall, but also several problems with likely significant impact on the overall performance as below:

- Does AWTT produce consistent data sets each time the experiment was performed?

It is not easy to obtain identical results even if the experiment is repeated with exactly the same steps and same start nodes. Due to the completely diversified and loose structure of the web, there are new web sites created and added to the web every hour. Similarly, old pages being removed increases the dead links. Therefore, it is unlikely to have followed the same traversing route or terminate at the same page. As a result, different data sets will be collected.

Besides, server down time, connection time out setting and Internet traffic condition varies from time to time. Therefore, detailed experimental reports *(as per attached in appendix (a) and (b))* were prepared to clearly record down the exact date and time when running the experiment. In addition, the pages at which the traversing stopped and the reasons, which it caused were also stated.

- Do significant differences in results exist within the same data sets if representative items and evaluation methodology change?

Feature extractions are definitely not limited to keyword based. There are alternatives, each with different characteristics and complexity (Barfourosh et al, 2002). Matching word vectors is one of the straightforward and most commonly used attempts for dealing with document related "questions" (Li, 1999).

On the other hand, precision and recall employed here have been well recognized for evaluating both classical and web information systems (Kobayashi and Takeda, 2000). Certainly, this evaluation method can also be altered based on individual interest. Kleinberg (1998) suggested that precision and recall for web system could be extended

to two different aspects: relevancy of results in the first page and the most information rich pages found i.e. authorities and hubs pages (section 2.1.2). All these no doubt have implication on the analysis results in certain degree if it applies to evaluate the current data sets.

6.0 CONCLUSION AND FUTURE DIRECTIONS

This chapter summarizes and concludes the dissertation. In addition it suggests some potential areas for further improvements to address the limitations in this work. Some future trends and research directions are also presented.

6.1 Summary and Conclusion

In this dissertation, a prototype was built, experimented, and finally evaluated. The primary goal is to conduct a preliminary investigation on the development of an automated traversing tool using elementary traversing approaches, specifically for collecting Malaysia web pages based on the web structure concept.

In particular, the prototype has implemented two fundamental traversing approaches: breadth-first and depth-first. Each approach has been tested to collect web pages started from 50 different nodes (URLs). These datasets were then assessed in respect to the precision and recall. One may argue that it is not necessary to consider breadth-first because the problem for finding Malaysia web pages is already a domain specific search. However, if viewed from other angle, finding web pages of Malaysia is not as small as for instance, finding "food" pages. The categories and the nature itself is broad enough to include both browsing and searching as defined in section 2.2.1.

The result from this investigation yields both positive and negative data, which support the current research work. The most comparable to the preliminary results shown here are those reported on traversing approaches and web structure (as described in more

detailed in chapter 5.0). A total of 100 collections (50 sets for each approach) or 8020 web pages were grabbed by AWTT. Based on the experimental results, the breadth-first traversal approach is less efficient way for collecting Malaysia web pages. A breadth-first traversal might gather a much larger collection of web pages but it is not able to collect more relevant pages. This could be due to the implementation in this work where breadth-first visited a randomly selected page in each level and subsequently traverse to the next level. In this case, the pages visited are always getting further from the starting page in terms of content relatedness.

On the other hand, the results obtained have clearly shown that depth-first collects pages that give higher relevancy percentage. However, it gathered a much smaller collection of web pages compared to breadth-first. Again, this may relate to the specific implementation here where depth-first tends to visit all pages in a level before proceed to the next level and hence forces the traversal drilling down to topically related pages. In comparison to breadth-first, depth-first often gets stuck in the process of traversal and more often reaches at deadlinks that terminate the process. In a broad-spectrum, depth-first is more likely to give a better overall distribution of URLs over the web, which may be important especially when a relatively small part of the web is to be retrieved. Nevertheless, the time taken for AWTT with the approaches of depth-first and breadth-first is not measured and compared, because of its close relationship with several uncontrollable factors including maximum connection time out set by different web servers, unpredictable downtime of web servers and inconsistent Internet traffic.

Additionally, the work from this dissertation can possibly be contributed to several areas. All components of AWTT prototype, which have been implemented in Object-oriented approach, can be treated as groundwork and reusable for the development of an operational actual web automated traversing system; the final and improved system hopefully may subdue the difficulty of finding relevant information on the web, more particularly Malaysia web pages. As for applicative scenario, making use of the current web structure rather than purely linguistics approach encourages a different perspective for collecting Malaysia web pages. Certainly, with more research on using this current web structure technology in a good manner may provide even better results. At the same time, the analysis result, which focuses on local web pages as dataset, has revealed that some Malaysia web pages development have been position in less competitive foothold from the context of web structure, most probably due to lack of linkages with the core portion of the WWW as a whole.

6.2 Limitations and Suggestions for Improvement

A number of weaknesses suggest some possible technical enhancements to the current AWTT. These disabilities can be observed from three areas.

As can be seen from the experimental results, for both approaches, the highest precision percentage achieved by both approaches is 20%. Currently, AWTT is a fully automated system without any user intervention throughout the process. For reasons of quality, this process also requires the human in the loop to input and manipulate. A component that captures users feed back can be incorporated into the traversal to construct

optimised traversal path to fetch back more relevant pages. With this mechanism in place, a list of web pages collected in the current level will be sorted in a personalized manner according to users' favorite web pages. The feedback loop causes AWTT to more accurately capture the features of Malaysia web pages. This problem itself may branch out to an interesting research individually (discussed in section 6.3 later).

A second area stems from an obvious limitation of AWTT as currently implemented. As pointed out by the dataset, the depth-first traversal of AWTT returned a relatively small collection of web pages because its crawling usually limited to 2-3 levels. There are clearly opportunities to increase the number of levels traversed by more appropriately handling exceptions such as obsolete links, non-html pages (i.e. pdf files, text files, scripting pages etc), connection time out etc. This is more a design and development consideration than a research topic, expanding the flexibility and strength for the current version of AWTT.

The third area of the current work involves implication of keeping a large and complex collection of traversing route in AWTT. As the issue of scalability arises, the current simple tables are unlikely to satisfy the needs. So applying a more proper database design might help to improve the efficiency for scheduling traversal process.

Besides, focus is placed only on the WWW pages rather than other applications, i.e. FTP, telnet, email hosts and servers etc.

6.3 Some Trends and Future Research Directions

This dissertation takes the development of AWTT and its experiment as a starting point to discover and collect solutions to automate the process of finding Malaysia web pages. Over the past few years, there are several important trends and future directions of some related research, which might have significant effect to the problem defined above.

Of special interest is the use of Artificial Intelligence techniques. Different algorithms and methods for machine learning are used in web information retrieval systems. Three main categories i.e. supervised learning, semi-supervised learning and unsupervised learning are usually implemented to improve the functionality of the systems (Bafourosh, 2002). These could provide new ways whereby the accuracy of pages fetched by AWTT could be improved, as describe in section 6.2.

Over the past few years, there is already a strong need for search engine technology to improve their traversing engine at the back end and have lately turned to agent technologies (Jansen, 1996). Likewise, domestic engines and many Malaysia enterprise portals that currently rely on user submission or manual insertion by web master to populate their database may employ agent technologies to more proactively search for local information from the web.

One typical application area noticed recently would be an emerging technology, which put isolated information into a meaningful context known as semantic web. As John et al (2003) pointed out "important information is often scattered across web and/or intranet resources but traditional search engines return ranked retrieval lists that offer little or no information on the semantic relationships among documents." As the current

WWW is slowly transforming to a new edition, for sustainable reason, AWTT must explore new search technology as well as innovative technologies for information processing and thus more selectively and effectively move across the semantic web.

Following several development trends and promising research described above, it is not uneasy to envisage that agent technology, artificial intelligence techniques and semantic web are the kind of directions for AWTT to embark on.