PRELIMINARY INVESTIGATION OF COLLECTING MALAYSIA WEB PAGES USING AUTOMATED TRAVERSING TOOL

Zaitun Abu Bakar
Department of Information Science
Faculty of Computer Science & IT
University of Malaya
50603 Kuala Lumpur, Malaysia
email: zab@um.edu.my

Tai Sock Yin
Department of Information Science
Faculty of Computer Science & IT
University of Malaya
50603 Kuala Lumpur, Malaysia
email: sytai@perdana.um.edu.my

ABSTRACT

Over the past few years, Malaysia web pages have gained popularity in the internet due to the increasingly strong demand of localized contents. Such huge document sets have introduced many new challenges to efficiently search among all other web pages. This study investigates the use of an automated traversing prototype that implements breadth-first and depth-first approaches to gather Malaysia web pages from the web. In the introduction, we describe the web structure and traversing approaches. Then, we discuss briefly on the experimental set-up that investigate the process of automating web traversal process to gather Malaysia web pages and compare the quality of information found using the two different traversing approaches i.e. breadth-first and depth-first. This is followed by a presentation of the results obtained and its analysis. Finally, the paper describes how the use of these traversal approaches can achieve different results.

Keywords: ***Web navigation, Traversal approaches, Search engines, Information retrieval, World Wide Web***

1.0 INTRODUCTION

As the size of World Wide Web (web) 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 [1], who also show an exponential growth in numbers of web pages, similar to the trends of web 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.

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 [2]. 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 [3].

In spite of FTP (File Transfer Protocol), telnet, emails, and web, 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 [2] 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 customised 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”. However, with the volume of information on the Internet growing exponentially to an astronomical figure, this 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 [2].

The objective of this paper is to discuss the investigation on gathering Malaysia web pages from the web using an automated traversing prototype that implement breadth-first and depth-first approaches. In Section 2.0, the paper will first introduce some compelling findings or research work on web structure and traversing approaches. And, in the subsequent section (Section 3.0), the paper will provide an overview of the experimental set up. Finally, in Section 4.0, some results will be presented and discussed before the paper is concluded in Section 5.0.

2.0 BACKGROUND

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 utilise the existing web structure to find pages, which discuss on Malaysia.

2.1 The World Wide 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 [4]. Internet is the core engine that provides the connectivity throughout the world but it is accessible via other application technologies such as World Wide Web, FTP, telnet, and emails. In this paper, the focus is placed in the web rather than other applications on the Internet i.e. FTP, telnet, email hosts and servers etc.

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 [5] that are based on the experiments on local and global properties of the web graph have shown a great deal of self-organisation. This study reveals 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 [6]. In this case, when web pages are randomly added, the web tends to organise 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. Therefore, the web can be alternatively seen to have arranged into “communities” based on number of links.

In normal circumstances, web page authors insert hyperlinks to pages that they are aware of because those sites are popular. Undoubtedly, this 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. 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 finding Malaysia web pages.

2.2 The Web Structure

A study conducted by A. Broder et al. [7], 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. 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 behavior of the web, and yielded insights into the complex organization of the web. The theory emphasises 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 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 categorised 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.

2.3 Automated Web Traversing

Regardless of manual or automated, eventually what do searching web pages mean? In general, searching can be loosely defined as a process for mapping users’ specified needs with the information available [2]. More specifically, the web search or Internet search, as described by Barfourosh et al. [8], 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 [9].

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.

The information structure of the web 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 [2]. Unlike the orderly world of conventional library collection, information on the web is chaotic, often not organised and jumbled up without clear boundary or separation. Mark Nelson [10] names this scenario as information anxiety - the overwhelming feeling one gets from having too much information or being unable to find or interpret data [11]. 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 organising such a huge collection. As Brian Pinkerton [12] states, “the World Wide Web is decentralised, dynamic and diverse; navigation is difficult and finding information can be a challenge” [12].

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 this paper, a prototype, known as Automated Web Traversing Tool (AWTT) was developed to implement the traversing concepts of breadth-first and depth-first. The use of these traversing approaches aims to take advantage of the existing structural topology of the web, which is the hypertext links structure. In general, the AWTT is used to investigate the automated process of browsing and searching for Malaysia web pages as well as to compare the quality of web pages found using these two approaches. Detailed descriptions on the development of AWTT are beyond the scope of this paper.

2.3.1 Breadth-first and Depth-first Traversal

In terms of breadth-first and depth-first approaches, page-based implementation was used here. This is dissimilar with the ordinary server-based implementation. Fig. 1 and Fig. 2 exemplify the details of these approaches.

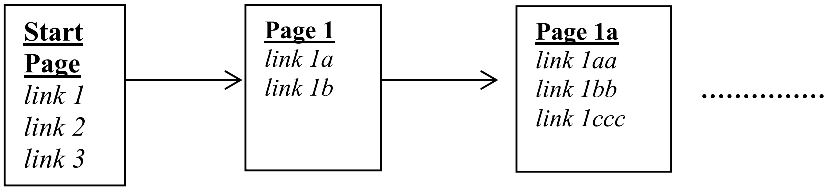


Fig. 1: Breadth-First Traversal of Page Based Implementation

From example in the Fig. 1, breadth-first traversal starts with visiting *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 on 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.

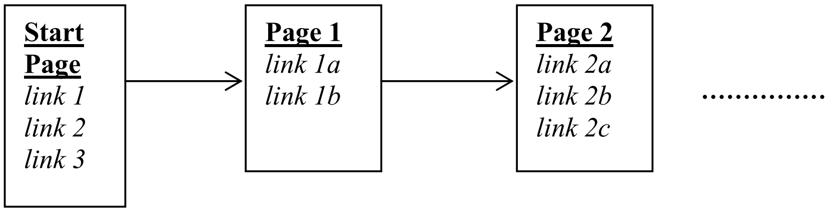


Fig. 2: Depth-First Traversal of Page Based Implementation

Similarly for depth-first traversal, no physical location of web pages in the server is considered for deciding the traversal path. From Fig. 2, 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.

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.

3.0 EXPERIMENTAL SET-UP

This section explains the experimental set-up on the data preparation process and some steps taken to conduct the experiment.

3.1 Objectives

There are two ultimate goals for this experiment: (1) investigate the process of automating web traversal process to gather Malaysia web pages and (2) compare the quality of information found using breadth-first and depth-first traversing approaches.

3.2 Data Collection

The data preparation process is mainly to deal with selecting suitable sample data sets to 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.

3.2.1 Start Nodes

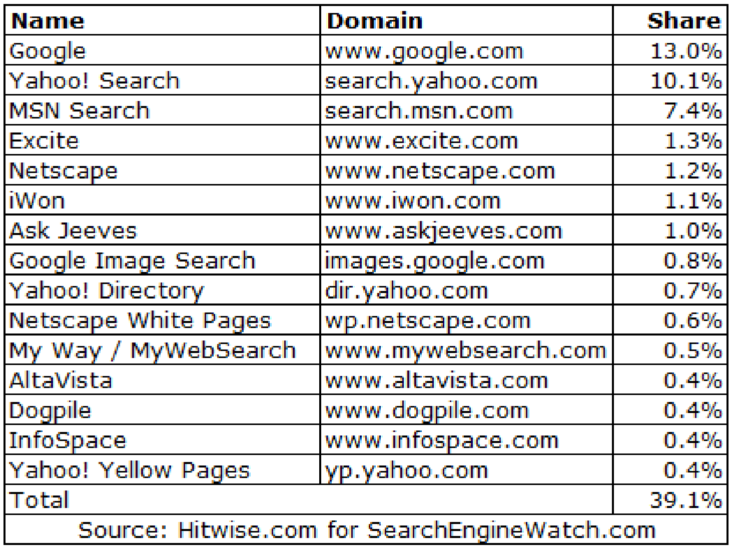
Despite a user input to manually indicate a start node, deciding a start node actually means identification of a good starting point. A good starting here can be generally seen as a node that has high probability to link with other relevant nodes in the same domain.

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 described in Section 2.2 above), 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, 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, 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 [13] dated 28th October 2003; there are several popular search engines, as shown in Table 1. 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.

Table 1: List of Most Popular Search Engines [13]



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.

There is one last search engine, which is not listed in Table 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 has definitely led to the increase of Malaysia web pages as revealed by the current ISP market and assesses future trends [14]. 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.

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. In view of this, Lycos was selected because it has an international database plus a more comprehensive and local-focused collection [15].

3.2.2 Size of Datasets

The target size of datasets for this experiment is about 10,000 Malaysia web pages. According to the latest statistics from MYNIC (The Malaysian Network Information Centre) [16], 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 [16]**.** In this case, the figure 10, 000 is indeed sufficient to give a fair view to this experiment.

3.2.3 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.

3.2.4 Assumptions

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 [17]. 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).

3.3 The Experiment Process

This section summarizes the experiment process. The data preparation stage starts with collecting Malaysia web pages from six major search engines, 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 dead links. Those results were eliminated and finally 50 web pages were selected as start nodes for the subsequent stage. The 50 start nodes were then 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.

4.0 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 Section 3.0. 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.

4.1 Analysis Methodology and Preprocessing

In this paper, precision and recall are used to measure the results generated by traversing approaches. 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.

4.1.1 Measuring Precision

Precision is one of the important aspects to be assessed in this work. 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 of 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 [2]. 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 [18].

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. 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 [19], 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

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 stop words [18] are eliminated. Then, Porter stemming algorithm [20] is used to stem all words extracted from web pages

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

4.1.2 Measuring 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. At here, 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 methods 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.

4.2 Results and Discussion

This section analyses the precision analysis results derived 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 Fig. 3, 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 but according to Tsunenori [21], 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 5.0 to improve this aspect of AWTT for future use.

Fig. 3 also shows 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 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.

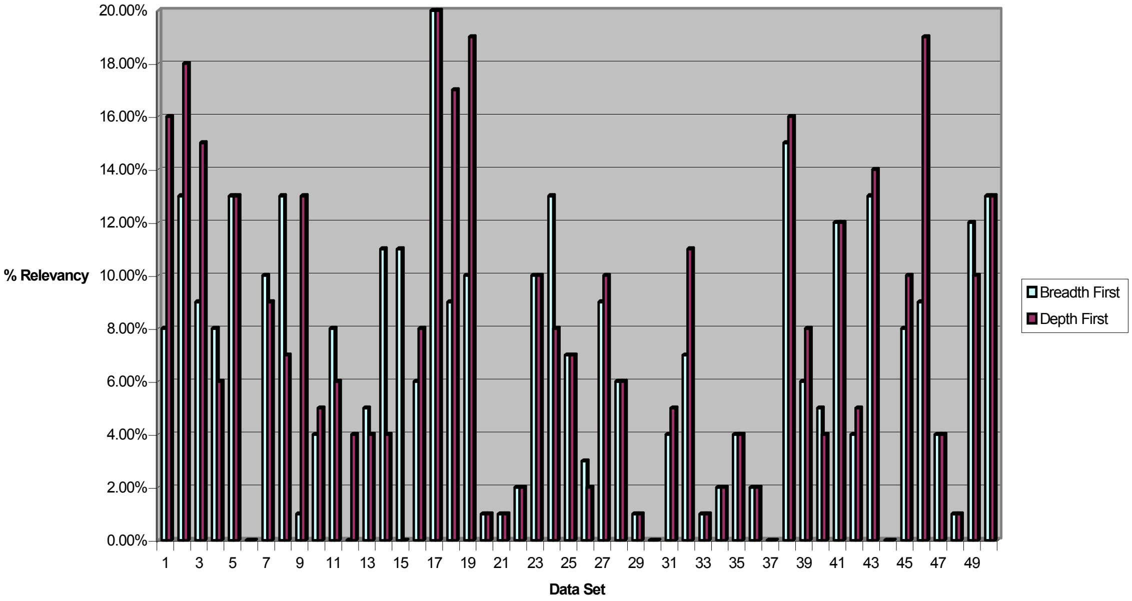


Fig. 3: Relevancy of Web Pages

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.

Fig. 4 shows an example of the changes of 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 [7] 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.

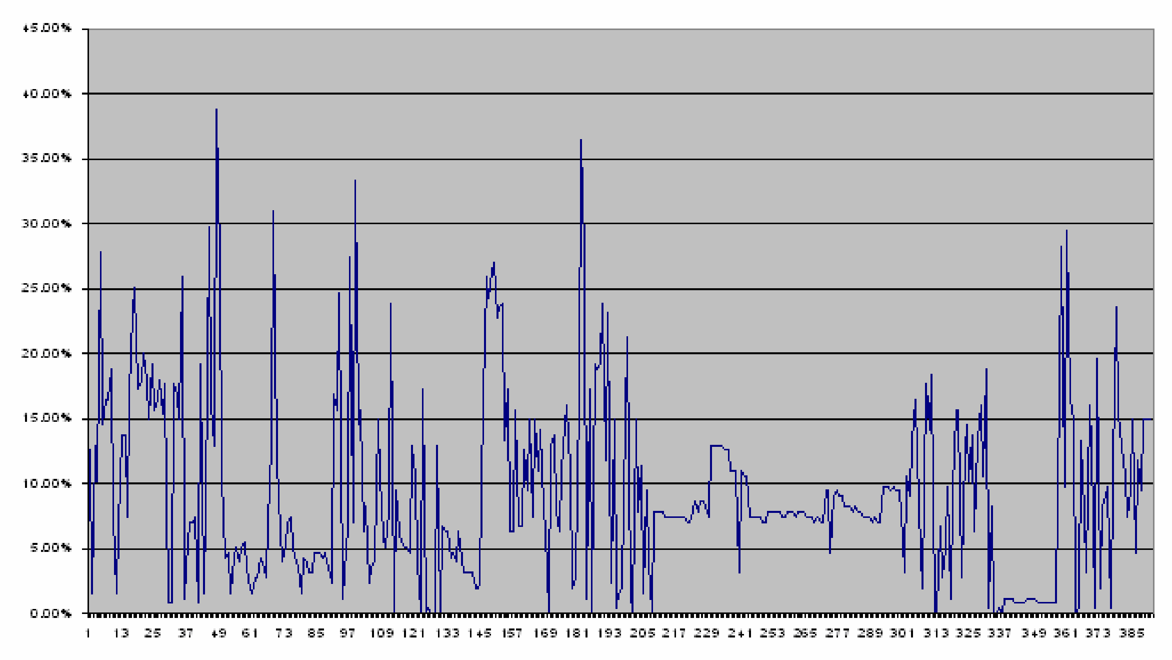


Fig. 4: Changes of Relevancy

Whereas from Table 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, 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 leads to the discovery of more new pages, which could be sitting at the origination or core portion of the web. As being identified, 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 [22]. This causes the amount of pages grabbed a lot more than depth-first as Table 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 study [26] have shown that depth of user visit within a website is 2.98, almost three pages as users usually 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). Therefore, 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. The study was also closely relevant to accessing the quality of web site.

In summary, different traversing approaches generated different traversal results. For example, 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 collected. Opposite to the breadth-first search, depth-first locates relevant but highly distributed information by assuming that relevant documents to a topic should be near each other in link structure.

Table 2: Total Number of Web Pages Collected

Start Node	Breadth-First	Depth-First	Start Node	Breadth-First	Depth-First
1	930	2	26	7	2
2	248	8	27	1	1
3	392	8	28	1	1
4	1000	2	29	6	2
5	1	1	30	1	1
6	5	0	31	1	4
7	211	2	32	5	4
8	1	2	33	15	1
9	4	2	34	1	1
10	5	2	35	1	1
11	110	2	36	2	2
12	1	1	37	1	1
13	32	2	38	20	2
14	915	2	39	4	1
15	960	0	40	36	4
16	934	8	41	507	7
17	1	1	42	5	4
18	846	2	43	517	4
19	20	2	44	0	0
20	2	2	45	2	1
21	2	2	46	117	3
22	2	2	47	1	1
23	3	3	48	2	2
24	6	2	49	5	3
25	9	5	50	1	1

4.3 Some Issues

In examining the 50 experimental datasets, 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.

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.

Besides, feature extractions are definitely not limited to keyword based. There are alternatives, each with different characteristics and complexity [8]. Matching word vectors is one of the straightforward and most commonly used attempts for dealing with document related "questions" [23].

On the other hand, precision and recall employed here have been well recognised for evaluating both classical and web information systems [24]. Certainly, this evaluation method can also be altered based on individual interest. Kleinberg [25] 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. All these no doubt have implication on the analysis results in certain degree if it applies to evaluate the current datasets.

5.0 CONCLUSIONS

In this paper, a prototype of automated traversing tool was built and experimented. 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 are broad enough to include both browsing and searching.

The results from this investigation yield both positive and negative data. 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 dead links 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.

Obviously, there are a number of weaknesses that require some technical enhancements. For reasons of quality, a component that captures users feed back can be incorporated into the traversal to construct optimised traversal path to fetch back more relevant pages. Besides, 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. 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. Further research on implementation of agent framework or artificial intelligence techniques might have significant effect to the problems defined above.

This work 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. The work can possibly be contributed to several areas. 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.

REFERENCES

[1] SIL (Summer Institute of Linguistics) (2000). *Ethnologue: Languages of the World*, 14th Edition. Available from: http://www.sil.org/ethnologue/countries/Asia.html. [Accessed 15th October 2003]

[2] G. J. Kowalski, and M. T. Maybury, *Information Storage and Retrieval Systems: Theory and Implementation*. 2nd Ed. Massachusetts: Kluwer Academic Publishers, 2000.

[3] S. William, *High-Speed Networks and Internets: Performance and Quality of Service.* 2nd Ed. New Jersey: Prentice Hall, 2000.

[4] P. Vincent, *Free Stuff from the World Wide Web.* Arizona: Coriolis Group Books, 1995.

[5] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Trawling the Web for Cyber Communities". *8th WWW Conference Proccedings,* Edinburgh, 1998, pp. 403-415.

[6] J. A. Kleinberg, “Authoritative Sources in Hyperlinked Environment”. *Journal of the ACM*, Vol. 46(2), 1998, pp. 212-235.

[7] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener, “Graph Structure in the Web”. *9th WWW Conference Proceedings*, Amsterdam, 2000, pp.309-320.

[8] A. A. Barfourosh, M. Nezhad., M. L. Anderson and D. Perlis, (2002*). Information Retrieval on the World Wide Web and Active Logic: A Survey and Problem Definition* [online].
Available from: citeseer.nj.nec.com/barfourosh02information.html. [Accessed 12th April 2003].

[9] M. J. Bates, *The Design of Browsing and Berrypicking Techniques for the On-line Search Interface.* Online Review, Vol. 13 (5), 1989, pp. 407-431.

[10] M. R. Nelson, (2001) *Being held hostage by information overload*.
Available from: http://www.acm.org/crossroads/xrds1-1/mnelson.html. [Accessed 18th July 2003]

[11] T. Nelson, *A File Structure for the Complex, the Changing, the Indeterminate*. 20th National Conference Proceedings, Baltimore, 1965, pp. 84-100.

[12] B. Pinkerton, (1994). *Finding What People Want: Experiences with Wbcrawler*.
Available from: http://www.thinkpink.com/bp/WebCrawler/WWW94.html. [Accessed 17th July 2003].

[13] D. Suvellan, (2003). Hitwise Search Engine Ratings.
Available from: http://www.searchenginewatch.com/reports/article.php/3099931.
[Accessed 30th October 2003].

[14] J. Calvert, (2003). *Internet Services: Malaysia, Gartner Report*.
Available from: http://www.gartner.com/. [Accessed 6th November 2003].

[15] Hoffman Agency (2000*). Singtel, Lycos launch Malaysia web sites.*
Available from: http://www.hoffman.com/newsgram/news_03_24_00.htm.
[Accessed 6th November 2003].

[16] MYNIC (2003*). MYNIC statistics*. Available from: http://www.mynic.net/. [Accessed 6th November 2003].

[17] D. Suvellan, (2003). *Search Engines Features Chart*.
Available from: http://www.searchenginewatch.com/facts/article.php/2155981.
[Accessed 6th November 2003].

[18] W. B. Frakes and R. Beaza-Yates, *Information Retrieval: Data Structures and Algorithms*, New Jersey: Prentice-Hall, 1992.

[19] Mauldin, Michael L. and John R. R. Leavitt (1994). *Web Agent-related Research at the Center for Machine Translation.* Available from: http fuzine.mt.cs.cmu.edu/mlm/signidr94.html. [Accessed 15th March 2003].

[20] M. F. Porter, *An Algorithm for Suffix Stripping*. Program, Vol. 14 (3), 1980, pp. 130-137.

[21] Tsunenori Ishioka, “Evaluation Criteria for Information Retrieval System”. *Transaction of the Institute of Electronics Information and Communication Engineers*, Vol. D-1(5), 2003.

[22] Shi Weisong, W. Randy, E. Collins, and K. Vijay, (2002). *Workload Characterization of a Personalized Web Site and Its Implications for Dynamic Content Caching*. New York University. Available from: www.cs.nyu.edu/csweb/Research/TechReports/TR2002-829/TR2002-829.ps.gz. [Accessed 10th September 2003].

[23] Li Wentian (1999). *Zipf's Law*, North Shore LIJ research Institute. Available from: http://linkage.rockefeller.edu/wli/zipf/ [Accessed 2nd June 2003].

[24] M. Kobayashi and K. Takeda, “Information Retrieval on the Web”. *Research Report, RT0347,* April 2000, Japan: IBM.

[25] J. A. Kleinberg, and S. Lawrence, (2001*). The Structure of the Web.* Science, Vol. 294, pp. 1849-1851.

[26] L. D. Catledge, and J. E. Pitkow, “Characterizing Browsing Strategies in the World-Wide Web”. *3rd WWW Conference Proceedings*, Darmstadt, 1994, pp. 1-9.

BIOGRAPHY

Zaitun Abu Bakar obtained her PhD from the University of Malaya and currently is an Associate Professor at the Faculty of Computer Science and Information Technology, University of Malaya. Her area of specialisation is Information Science and she teaches Enterprise Resource Planning, Business Process Reengineering and Knowledge Management at the postgraduate level. Her research interest includes e-government, e-learning, Problem Based Learning, Workflow Systems and Women in ICT.

Tai Sock Yin holds a BSC (Hons) in Computing from the University of Portsmouth (U.K). She is currently pursuing her Masters (Computer Science) at the University of Malaya. Her current research interests include web information retrieval and agent technologies.