# Website Content Extraction Using Web Structure Analysis

by

Nor Hayati Binti Daraham

Dissertation submitted in partial fulfillment of
the requirement for the
Bachelor of Technology (Hons)
(Information Communication Technology)

DECEMBER 2005

Universiti Teknologi PETRONAS
Bandar Seri Iskandar
31750 Tronoh
Perak Darul Redzuan

1) web portals

# CERTIFICATION OF APPROVAL

## Website Content Extraction Using Web Structure Analysis

by

Nor Hayati Bini Daraham

A project dissertation submitted to the
Information Technology Programme
Universiti Teknologi PETRONAS
in partial fulfilment of the requirement for the
BACHELOR OF TECHNOLOGY (Hons)
(INFORMATION COMMUNICATION TECHNOLOGY)

Approved by,

(Miss Shakirah Mohd.Taib)

UNIVERSITI TEKNOLOGI PETRONAS
TRONOH, PERAK
December 2005

# CERTIFICATION OF ORIGINALITY

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

Norhayati

NOR HAYATI BINTI DARAHAM

# ABSTRACT

The Web poses itself as the largest data repository ever available in the history of humankind. Major efforts have been made in order to provide efficient to relevant information within huge repository of data. Although several techniques have been developed to the problem of Web data extraction, their use is still not spread, mostly because of the need for high human intervention and the low quality of the extraction results. For this project a domain-oriented approach to Web data extraction and discuss it application to extracting news from Web Sites. It will use the abstraction method to identify important sections in a web document. The relevance information will be taken account and will be highlighted in order to develop a focused web content output. The fact-finding and data about the project are gathered from various sources such as internet, and books. The methodology used is a Waterfall Model that involves several phases which are Planning, Analysis, Design and Implementation. The result of this project is the display and review of web content extraction and how it being currently being developed which the goals is to give more usability and easiness toward web users.

# ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious, the Most Merciful.

First of all, I would like to express my greater gratitude to Allah the Almighty, for His help and support during the course of life and the moment of truth.

Many people have contributed to the successful of my Final Year Project. I would like to thank my respected supervisor, Miss Shakirah Mohd. Taib for the encouragement, guidenance, insight and her strong support for this project. A special thanks also to the internal evaluators Mr. Abdullah Sani and Mr. Izzatdin for their comments during the first evaluation of my final year project.

I remain deeply indebted to my parents, siblings and relatives for their morale support and motivation throughout the journey in completing this final year project.

Last but not least, special thanks to my friends Fazilah, Nuraimi, Arnida, , Hazleen, and Izatun and other course mates for their critics but their ideas and enthusiasm nurtured me during the preparation of this project.

# TABLE OF CONTENT

CHAPTER 3: METHODOLOGY . . . . . . .

CHAPTER 4: RESULT AND DISCUSSION . . . . .

CHAPER 5: CONCLUSION AND RECOMMENCATION . . .

## LIST OF FIGURES

## LIST OF TABLE

# CHAPTER 1
# INTRODUCTION

## 1. BACKGROUND OF STUDY

The Internet contains more and more Web paged with dynamic and frequently updated data. If a search engine provides useful help for users to identify relevant information, it cannot be used to "understand" the semantics of the results and to obtain the reliable data. This is mainly due to the lack precision and standard formalism in presenting data and because HTML is a formatting language. In addition, current search engines are more focused on static data on the web rather than dynamic data that constantly change, such as weather forecasts, and stock exchange information. On the other hand, such data is more required by automated processes such as software agents.

Users are spending more and more time on the Internet in today's world of online shopping and banking; meanwhile, web pages are getting more complex in design and content. Web pages are cluttered with guides and menus attempting to improve the user efficiency, but they often end up distracting from the actual content of interest. These "features" may include script and flash driven animation, menus pop-up ads, obstructive banner advertisements, unnecessary images, or links scattered around the screen. In order to gain the desired information, an abstraction of layout structure can be done and it makes the interpretation of web layout structure become easier.

Content Extraction (CE) identifies parts of the document for extraction. Nowadays, it is more going into extraction of semantically information. The observation concerns "Extracting". Extracting on the Web is systematically searching through one or more sites in order to locate information of interest at those sites.

## 1.1 PROBLEM STATEMENT

### 1.1.1 Problem Identification

Web pages are often cluttered with distracting features around the body of an article that distract the user from the actual content they're interested in. These "features" may include pop-up ads, flashy banner advertisements, unnecessary images, or links scattered around the screen. Navigation features like menus and quick links often interrupt a user's workflow.

The problem that arise from the news web pages are as follows:

1. The users can't focus on their desired goals which are, to find the exact content that they really want during the search process in the internet.
2. Most of the user need to focus on the information that fits on their interest only and some unsuitable content have to be ignored or filtered them out.
3. The knowledge of web structure and contents is required so that the desired information can be easily extracted.

### 1.1.2 Significant of the project

There some benefits that users will get when using this application One of the motivations is to present the data to the user in a more focused type of output. When the user choose certain domain web that they desired, the user able to gain more concentration towards the data that the read from the web. This is due to the ability of the application to highlight the initial data that related to the right content. The application will be one of the reasons the user can gain to their information retrieval more easily and convenient.

## 1.2 OBJECTIVE

This project comprises three main objectives:

1. To give the ability for users to get the real information that they desired.
2. To develop a system that will classify the important output content from web pages.
3. To analyze the effectiveness of structure abstraction in order to extract web content.

## 1.3 SCOPE OF STUDY

The project will involve with choosing one domain which are news. Several web pages will be gathered and will consist of many type of web structure presentation. In order to extract the desired data that really related to news domain that has been selected, a set of rules will be used in order to classify the real content while the other irrelevance content will be excluded from being highlight.

The application that will be involved for this project will be Microsoft Visual C++. A code generation will be build in order to make the content of web pages can be filtered and categorize as a related and relevance content for the news domain or not.

### 1.3.1 The Relevancy of the Project

Due to the growing number of online accessible web documents and the density of content's organization Web content's extraction and analysis become more challenging. The web design presentation is being displayed in variety web styles using numerous available web editors. It is being reported that web content extraction and analysis is very important process in information retrieval system, web classification and monitoring system. For web classification, there are many techniques applied based on keyword categorization. In this project, the system will able to give the ability to the user of having their desired data. The cluster information display can be handle and the web users can be more focused toward their desired information.

### 1.3.2 The Feasibility of the Project within the Scope and Time Frame.

The time given to complete this project is 14 weeks. Within time given, it is important for the author to complete the project successfully. The scope of the project is clarified first to make the objectives that the author want to achieve can be a reality. The system is hopefully to be finished on time within the time frame.

Secondly, the web user can gain the information that they desired faster. This is because the application will able to highlight the related information, the user just need to find and read the chosen data and able to extract the information that they want.

# CHAPTER 2
# LITERATURE REVIEW

## 2 THE IMPORTANCE OF WEB SITE

According to James Hobart (1990) the World Wide Web (WWW) has become a very popular means for publishing information. A large number of information repositories (Websites) already exist and new ones are being created at a very rapid rate. Most of the pages on WWW repositories provide elements that allow users (readers) to interact with them. Thus, the people designing pages for the WWW are actually designing user interfaces.

### 2.1 What is Web Content Extraction

Content Extraction (CE) identifies parts of the document for extraction. Nowadays, it is more going into extraction of semantically information. The observation concerns "Extracting". Extracting on the Web is systematically searching through one or more sites in order to locate information of interest at those sites.

### 2.2 Effort in Web Content

Major efforts have been made in order to provide efficient access to relevant information within this huge repository. At least two broad views of the problem have evolved recently. The first one, characterized by the unstructured view of data, has developed breakthrough technologies (such as Web engines) based on information retrieval, (Baeza-Yates, 1999) methods, which have been used in many successful commercial products. The second one, characterized by the structured or semi-structured view of data borrows techniques from the database area to provide the means to effectively managing the data available on the web, (Florecu, 1998). Thus, several techniques have been adapted to the problem of extracting data from the Web for further processing (querying, integration, mediation, etc), (Laender, 2003). However, these techniques are still not spread as the information retrieval based ones.

Devising generic methods for extracting Web data is a complex task, since Web is very heterogeneous and there are no rigid guidelines on how to build HTML pages, and how to declare the implicit structure of the Web pages. Thus in order to develop effective methods for extracting Web data in precise, it usually required to take into account specific characteristics of the domain interest.

## 2.3 Approaches to solve Web Content Extraction Problem

There are mainly three approaches to deal with this problem of data extraction from the web pages.

The first approach relies on natural language processing (NLP). It is known that current NLP is not accurate and powerful enough to recognize the contents of unrestricted web pages. Therefore, this approach has only been used in some limited areas.

The second approach tries to associate a web page with some semantic markers (or tags) when it is created. For example, one may use personalized markers. The limitations of such approach are well known, since the markers are personalized, they can hardly be generalized (Atzeni, 1997). Currently, an initiative of Semantic Web, W3C-Semantic Web (1997) is geared towards the creation of a web structure that more readily recognize the semantics of Web pages. The method currently under investigation consists in defining a general ontology of meta data on semantic contents. However, few actual Web pages use such markers.

There is a third manner to solve the problem. As the original data are structured in different ways, it is more suitable to structure them according to a common model that is independent of the information resources. Thus, extracting and combining data from different sources will be mush easier and more reliable.

## 2.4 Related Work

Several research groups have focused on the problem of extracting structured data from HTML documents. Much of the research is in the context of a database system, and the focus is on wrappers that translate a database query to a Web request and parse the resulting HTML page.

A group of researchers (Rahman et al., 2001) propose techniques that use structural analysis, contextual analysis, and summarization. The structure of an HTML document is first analyzed and then decomposed into smaller subsections. The content of the individual sections is then extracted and summarized. While the paper describes prerequisites for content extraction, it doesn't propose methods to do so. The solution is meant for constrained devices like cell phones, but implementation is more of a one-size-fits, as the technique is not adjustable; therefore an administrator has low flexibility retrieving removed content. Limitations of their work include the lack of ways to generate a good summary from documents that have multiple main themes, have specific constructs such as bullets and lists, cross comparing section headings with text, and detecting relationship among sections for safe merging. The screenshot of web page example that being shown from the projects is as Figure 2.1:

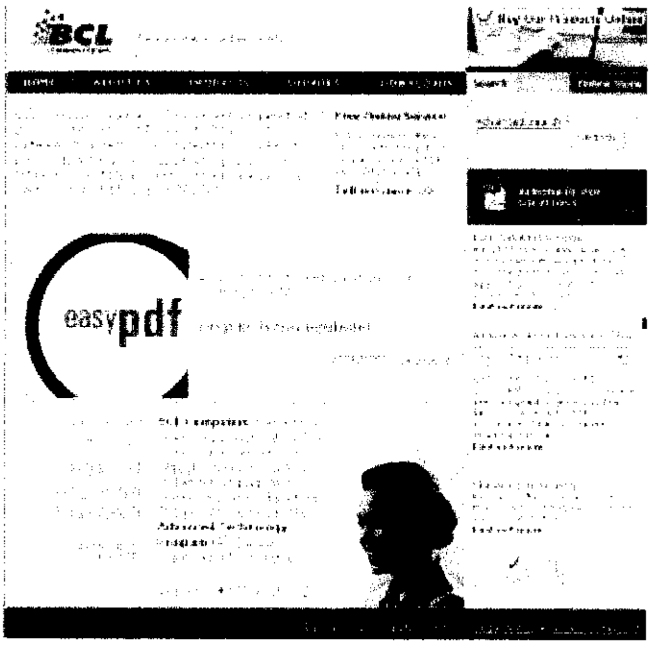
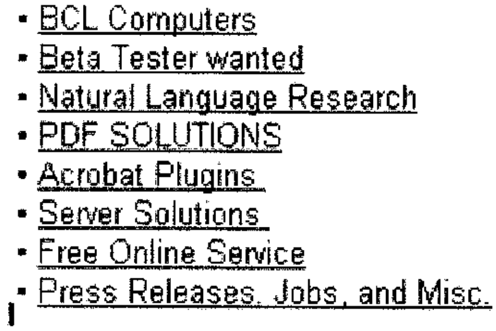


**Figure 2.1 Sample web pages**



- BCL Computers
- Beta Tester wanted
- Natural Language Research
- PDF SOLUTIONS
- Acrobat Plugins
- Server Solutions
- Free Online Service
- Press Releases, Jobs, and Misc.

**Figure 2.2 Summarized output**

Another group of researchers (Finn et al., 2001), discuss methods for content extraction from “single-article” sources, where content is presumed to be in a single body. The algorithm tokenizes a page into either words or tags, the page is then sectioned into three contiguous regions, placing boundaries to partition the document such that most tags are placed into outside regions and word tokens into the center region. This approach works well for single-body documents, but didn’t produce good results for multi-body documents, i.e., where content is segmented into multiple smaller pieces, common on Web logs (“blogs”) like Slashdot (http://slashdot.org).

Some researchers (Kaasinen et al., 2000), discuss methods to divide a web page into individual units likened to cards in a deck. A web page is divided into a series of hierarchical "cards" that are placed into a "deck". This deck of cards is presented to the user one card at a time for easy browsing. The paper also suggests a simple conversion of HTML content to Wireless Markup Language (WML), resulting in the removal of simple information such as images and bitmaps from the web page so that scrolling is minimized for small displays. While this reduction has advantages, the method proposed works with flat HTML files rather than a tree based approach that created problems in differentiating layout tables from other tables. The problem with the deck-of-cards model is that it relies on splitting a page into tiny sections that can then be browsed as windows. However, that means it is up to the user to determine on which cards the actual contents are located.

A similar approach to the deck of cards was proposed by (Chen et al).They used the same concept except in their case using the Document Object Model (DOM) tree for organizing and dividing up the document. They proposed to show an overview of the desired page so users can select the portion of the page they truly interested in. When selected, that portion of the page is zoomed into full view. One of their key insight is that their overview page is actually a collection of semantic blocks that the original page has been broken up into, each one color coded to show the different blocks to the user. This, very nicely, provides the user with a table of contents from which to select the desired section. While this is an excellent idea, it still involves the user clicking on the block of choice, and then going back and forth between the overview and the full view. Since improving a user's workflow is one of the target applications, their goal is to extract and provide the relevant content right away instead of making the user go back and forth between the content and the table of contents.

Other approaches(Suhit, Kaiser, et, 2003), that being displayed is based on DOM-based also, they employs multiple extensible techniques that incorporate the advantages of the previous work on content extraction. In order to analyze a web page for content extraction, the page is first passed through an HTML parser that creates a DOM tree representation of the web page. This process accomplished the steps of structural analysis and structural decomposition in techniques done by Rahman, Buyukkokten and Kaasinen. Their approach depending on the type and complexity of the web page, the content extraction suite can produce a wide variety of output. The algorithm performs well on pages with large blocks of text such as news articles and mid-size to long informational passages. Most navigational bars and extraneous elements of web pages such as advertisements and side panels were removed or reduced in size, where figures 2.3 and 2.4 below show how the output display will be seen as an example.

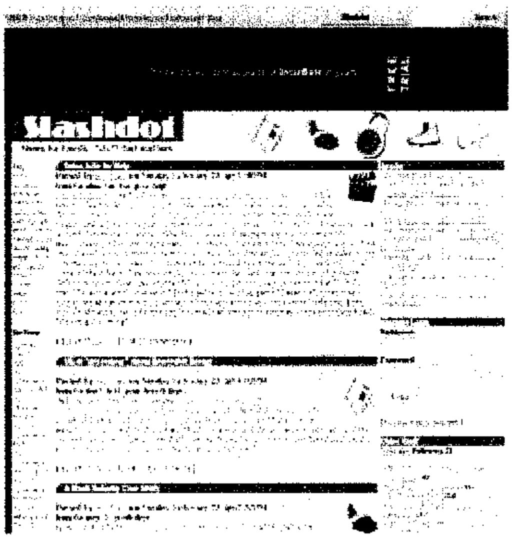


**Figure 2.3 – Before**

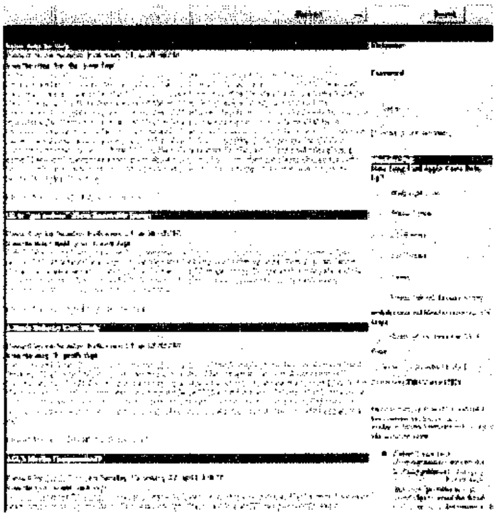


**Figure 2.4 – After**

## 2.5 Structure and criteria for Optimal Web Design

The organization of information within websites is vital o its overall usefulness. In fact, a study by (Morkes and Nielsen,1997) found that their experimental websites scored higher in usability when text was written concisely (58%), easily scan able (47%), written in an objective instead of a promotional style (27%), than web pages in their control condition.

That is, viewers tend to move quickly from page to page. Instead they usually scan for information that is of direct interests them. Accordingly, it is suggested that text should be very succinct, include only one key idea per paragraph, use highlighted keyword or phrases, and use bulleted lists when possible.

### 2.5.1 Web Views

Users often miss important pieces of information simply because it is not seen. This often occurs because they forget or are unwilling to scroll in a particular directly especially horizontally, and thus do not see the information that is located outside of the primary viewing area. To reduce this problem, important website information should always fit within the typical horizontal viewing area of he screen. To do this, the rule is still o design for lower resolution settings. According to real-time analysis of Web surfers by MyComputer.com, 800 x 600 currently is the most frequently used computer screen resolution.

The actual usable size to avoid any scrolling at this resolution is 595 x 295 pixels and the safe width for printing at this resolution is 535 pixels. Most users however have their resolution set at 800 x 600 (31%). To avoid scrolling here, the usable size is 750 x 425 pixels. A compromise would be to place the most important information within areas that are visible at lower resolution settings, while placing less important information in areas visible at higher resolution settings.

In addition, when users do scroll, they may not see the information because it is place in the typically low information- priority area, such as he bottom of a page (Nielsen, 1999) or placed in an area where users typically would not expect it to be placed.

### 2.5.2 Web Layout

Much has been said about the design process of websites, such as establishing the proper mood or "feels" to create user interest or even excitement with the site. This is a very important concern, but ultimately users tend to be far more satisfied and stay with websites that are designed for their use in mind (Tedeschi, 1999). Considering this, three core principles concerning interface design are presented:

1) Keep the interface simple- To quote Mies van der Rohe, "less is more." Organize the interface by reducing un-needed visual elements as much as possible. That is, remove all unnecessary visual "noise". This will make the important objects that are here stand out even more. Moreover, as Edward Tufte stated, "it is not how much space there is, but rather how it is used and it is not how much information there is, but rather how effectively it is organized" (Tufte, 1990).

2) Make action-objects visible- According to Donald Norman, a design should
    a) Make it easy to determine what actions are possible at any moment.
    b) Make things visible, including the conceptual model of the system, the alternative actions, and the results of actions
    c) Make it easy to evaluate the current state of the system. On a web interface, one of the chief mechanisms to do this is the proper use of perceived affordances (Norman, 1988).

    An affordance refers to the "properties of the thing, primarily those fundamental properties that determine just how the thing could possibly be used". Affordance provides us with clues as to the operations of things. More importantly for interfaces, however, are perceived affordances that provide visual feedback that advertise affordances. For example, a link button may be perceived to afford clicking because of its

"3-D' or 'raised' appearance. Consequently, it is often helpful to give link buttons the physical appearance of a button, or any object that affords clicking, in order for them to be seen as a button to be clicked (Norman, 1988). Thus, it is important to make navigation button look like they should be clicked as well as as follow the convention of underlining links when they are text-based links. Conversely, non- navigation objects should not look like they could be clicked in order not to 'trick' the user into thinking they are links.

Generally, button serve as primary object for initiating actions, such as submitting or confirming information. Buttons also can act as the primary link for movement to other web pages, usually the same website. When this occurs, text-based links often serve as a less important, secondary or supplemental link for the buttons. Normally, however, text-based links are the primary link o other internal web pages.

Moreover, physical appearance of objects such as icons can significantly affect navigational performance. For example, (Rogers, 1987) found that icons with abstract but simple symbols that represented concrete objects resulted in the fewest number of errors and requests for help. In addition, (Byrne, 1993) found that large and simple icons outperformed complex ones by a significant margin. Byrne suggests that icons need to be simple, large, and easy to discriminate in order to be effective. Complex icons tend to clutter the screen with unnecessary information. Moreover, (Norman, 1985) suggests that icons are the best used to represent graphic tools and objects. Verbal labels, such as "to save" are best for formal commands.

3) Balance and unify the interface – Balance and unity has always been a key component in good design. Humans on a preconscious level seek structure in the things they see. If there is no intentional structure, we will impose our own. Seeking the appropriate balance among things, as well as unifying those things that are related will generate structures that are not only pleasing to the eye, but will make the interface more understandable (Mullet & Sano, 1995), for an excellent discussion on design and visual interfaces). Empirical studies have supported that claim by finding that the position within a plane as well as size and contrast to be one of the most perceptually important variables in visual search tasks (Cleveland, 1985).

One of the fundamental concepts of balance is the notion of the golden Section. The Golden Section is a ration of a rectangle in which the smaller side o the larger is the same as that of the larger to the sum of both which is a ration of approximately 0.618 to 1.000 or a standard 8.5 x 11 page. Examples of the Golden Section are almost ubiquitous in art as well as in nature the Parthenon to a nautilus shell. A web page that structures its graphical layout according to this ration will look more appealing and will have a greater impact than other ratios, such as a ratio of 1 to 1.

Also when placing several objects on a web page, one should take into account the "visual mass" of these objects neither its size nor presence. For example, ideally the placement of objects should be positioned in the same way as balancing solid objects on a fulcrum. That is a larger object should be placed closer to the center of the screen to offset the smaller object(s). This will create equilibrium between the objects, and will be more appealing (Tufte, 1990).

The unity of the interface is important because it has the potential to link concepts and objects together hat belong together. For example, (Wickens, 1986) compatibility of proximity principle states that the necessitate mental integration of information should be in close proximity. However, tasks that require focused attention on specific variables will be harmed by this close proximity. This can be applied to things such as the organization of links. For instances, care should be made to group links that belong together, as well as separate those that do not belong.

(Ngo and Byrne, 2001) have taken this notion several steps further by identifying characteristics that define an aesthetically appealing interface. Of the 14 characteristics identified, balance, equilibrium stability, and sequence **(shown in Figure 2.5 until 2.7)** scored high in aesthetic correspondence.

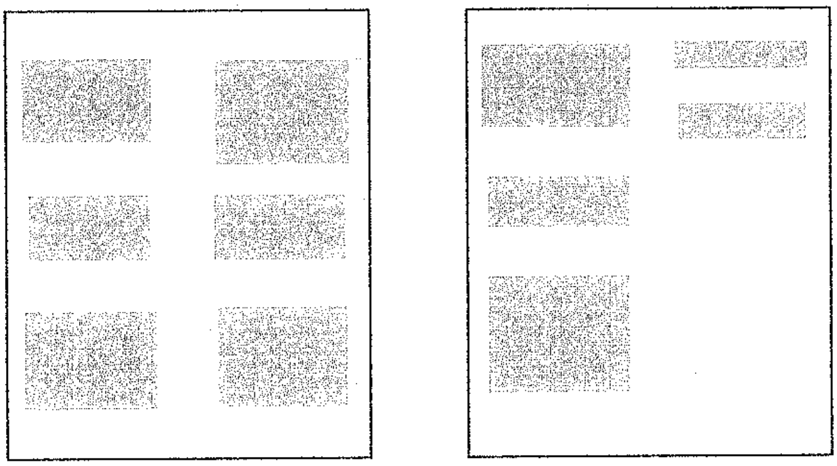


**Figure 2.5(a)** **Figure 2.5(b)**

Figure 2.5 is an example of well balanced (a) and poorly balanced (b) interface.

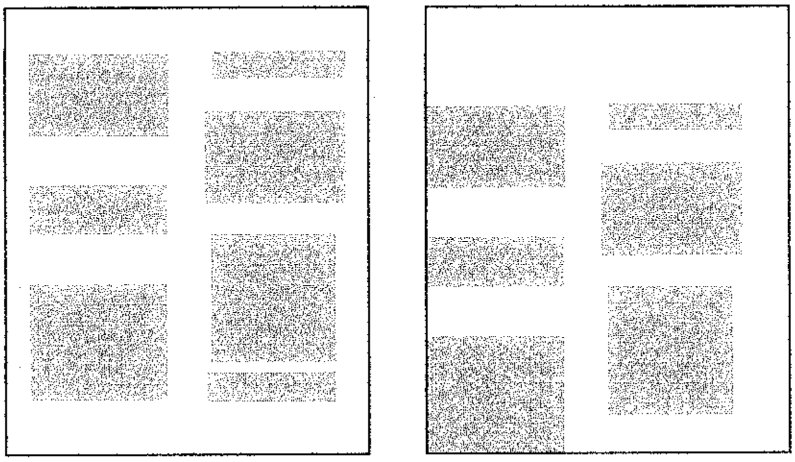


**Figure 2.6(a)** **Figure2.6 (b)**

Figure 2.6 is an example of an interface with stable (a) and unstable (b) screen. Equilibrium consists of the general centering of the interface itself to make it a stable arrangement.

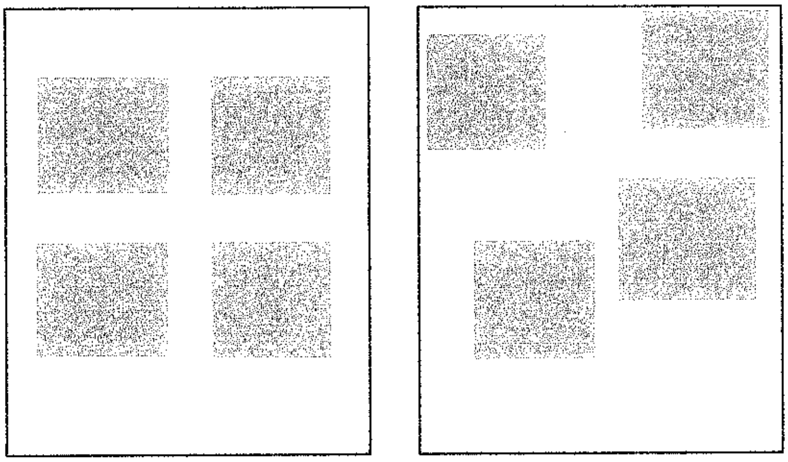


**Figure 2.7(a)** **Figure 2.7(b)**

Figure 2.7 is an example of a interface with a sequential (a) and non-sequential (b) screen.

Measuring subjective difference between well and poorly balanced interfaces, (Brady and Philips, 2002) found no statistical differences in user satisfaction, suggesting that user satisfaction is related more to successful navigation than aesthetic appearance. However, both (Brady and Philips, 1997) indicated that participants did perceived aesthetically pleasing sites as having a higher degree of usability.

Be aware of Fitt's Law- Formally, it states that pointing time is a function of the distance and the width of a target (Fitts, 1954). Generally speaking, it states that smaller and farther away an object is, the longer it will take to the reach that objects. Several researchers has argued that important button should be placed on the right side of the screen because the mouse arrow pointer is usually resting next to the scroll box, and thus it would take less time to click he object. However, what is important here is that knowledge that if the farther apart they are, the longer it will take to click them.

### 2.5.3 Web Structure and Navigation

People often become lost within the structure. In fact, 58 % of users will make two or more navigational errors while searching for information (Forsythe, 1996) and 66.8% of users have stated that one of the greatest problems about the Web is "not being able to find the information that are looked for". Generally here are four major reasons for is occurrence (Foss, 1989):

First difficulty is disorientation or "lost-in-hypertext problems, which arises from and unfamiliarity with the structure or conceptual organization of the site. The, users have difficulty deciding which node (which is typically one web page) to view next because they ere unable to visualize where the information they are looking for could be. The decision concerning which node to view next first involves understanding one's current location within he site, hen selecting the proper route. However, users may not even know their current location within as site.

A proper way to reduce this problem is to organize to problem is to organize the site according to the typical users' mental model of how a site would be organized. This can be done by having representative users sort cards into several categorical piles in which each card represents the information at would be placed on the actual website. Each pile should indicate the information that would be clustered within each category and subcategory. This would give the designer knowledge on how users mentally organize the structure of a particular side technique that uses this method discussed in Usability news (Bernard, 2000).

In addition, the placement of submenu titles may also help reduce disorientation. For example, (Gray, 1986) found that of the navigation errors made within hierarchy, 40% of them were in the third and fourth levels with submenu titles. Without submenu titles, 59% of the errors were made in the third and fourth levels. Moreover, according to (Bransford and Johnson, 1972), participants who have read passages and had higher comprehension than participants who did not have passages with tiles.

The use of navigational aids such as color code and consistent logos and banners should also reduce disorientation and the use of the "bread crumb" navigation techniques may help in reducing the disorientation problem as well.

The second difficulty is the embedded digression problem. This occurs when users pursue digressive paths within websites and lose their place or forget to return to their original document. This can be lessened by reducing the number of links embedded in text by placing them instead at the end or on the side of the document. However, (Knoved and Shneiderman, 1986) found that users preferred and were accurate in answering information using embedded links than explicit grouping of links outside the text. Yet, they also stated that embedded links could be disruptive in that the user "may be inclined to examine a particular subject or subjects in detail without first getting an appreciation of the overall context".

A recent study by (Bernard, Hull, & Drake, 2001) examined the effects of embedding associative links with a document, as well as placing them at the bottom, at the top-left, and left, at the same height in which they correspond with the document. No significant differences between the four link arrangements were detected in terms of search accuracy, time, or efficiency. However, there were significant subjective differences between the links arrangements favoring the embedded links. That is, participants indicated that they believed that embedding he links within the document made it easier to navigate, more easily recognize key information, promoted comprehension, and was easier to follow the main idea of the passages while searching for specific information. Moreover, participants significantly preferred the embedded link arrangement to the other arrangements. Conversely, placing links at the bottom of a document was perceived as being he least navigable arrangement, and was consequently least preferred. Thus, while embedded digression may be a problem for some users, this should be weighed against the subjective perceptions that favor the embedded link arrangement.

The third difficulty is the "art museum" problem. This refers to he lack of memory for the navigational details of a significant part of the site because he viewer is overwhelmed by he sheer amount of information. For instance, as when patron visiting a museum cannot hope to remember the details of all the artwork because of their great number, a large number an variation of navigational information such as the various node they have visited may consequently overwhelm he user. This often can have the effect of reducing a person's recall of the pages they have visited.

This can be lessened by reducing the amount of information presented at one time and properly organizing the navigational structure of the website. For example, in a study comparing three types of structures: pure hierarchical o the web pages at one level can only access by a web page directly above or below it, nonlinear which means links could be connected to any number of other web pages on the sites, and mixed design that is hierarchical structure with cross referential links, researchers found that participants recalled more information with the mixed design. The pure hierarchical structure was found to be too restrictive and the nonlinear design presented too much information at one time (McDonald & Stevenson, 1998). Thus, sites should present only the amount of links that are necessary for navigation- superfluous links will increase the probability that he users will be confused and disoriented. Additional support for this conclusion can be derived from the Hick-Hayman law, which generally states that the greater the number of options, the longer it takes to find the appropriate one because of greater uncertainty.

Other aids that are beneficial to navigation are the use of sitemaps. Sitemaps may, if done properly, present the structure of a site in a more cognitively manageable way by showing a site's main structure and the various links to that structure.

The fourth difficulty may be the structure itself. That is, it is generally found that people make fewer mistakes if the hierarchical structure. That is, it is generally found that people make fewer mistakes if the hierarchical structure of the site is broader rather than deeper.

In fact, research has generally found that ideally all information should be placed within three hierarchical levels from the initial homepage of the site. Specifically, the more levels users have to make in order to get the information hey want, the less chance they will find this information. For instance, in placing hyperlinks on a web page, (Larson and Czerwinski, 1998) point out that the moderate level of breadth is optimal if it is preceded by a well-organized layout. In their study, they reported that a two-level site beginning with 16 sequential links on the first level, then 32 links on the other produced reliably faster searches of information and produced less confusion than a three-level site with eight sequential links in all three levels. The reasoning here is that the deeper the levels, the more a user has to rely on short-term memory. Deeper level sites also have more general and consequently vaguer links descriptions at he top level, which makes it more difficult for users to figure out and remember the correct pats to a target.

However for sites that must have deeper structures by 4 or more levels, Norman and Chin (1998) found in their study of different menu tree structures hat users browsing for specific information will find this information faster if the structure that is concave (breadth of 8 x 2 x 2 x 8 pages). That is, it should be broad at the top level and at the lowest or 'base' level, while the interior of the web structure should have a narrower level of breadth (see Figure 2.8 below). They argue that a broad top level gives the user enough specific information to formulate an idea as to the correct path to take, while concentrating much of the information and the choices at the base level will help the user find that specific item. A narrower breadth interior will, in turn, reduce the likelihood of getting loss within the site because the user will have fewer choices, and consequently less chances of being disoriented.

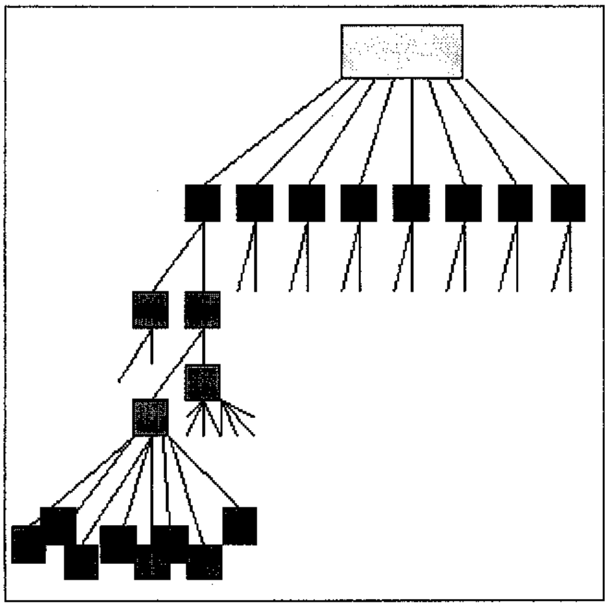


**Figure 2.8: Concave (8 x 2 x 2 x 8) menu tree**

As discussed by (Bernard, 2002), depth alone may not be the sole, or even the greatest determine in predicting search performance. In fact, as was shown, the shape of a hypertext structure had at least as much to do with search efficiency that its depth. Indeed, a (4 x 4 x 4 x 4) structure was found to be not only less efficient than hypertext shapes of the same depth for example, a (6 x 2 x 2 x 12) structure. As discussed, much has been said about hypertext depth, in that the greater the depth, the less informational efficient the structure should be for example, (Jacko & Salvendy, 1996: Snowberry, 1983). However, what seems to be occurring is that the search efficiency is at least in part, determined by the properties related to the overall shape of the hypertext structure. These properties, then, act to either help facilitate or impede hypertext efficiency by altering the general complexity of the structure. Accordingly, having an inefficient shape will decrease a hypertext's search efficiency.

Consequently, the goal should always be to reduce the complexity of the site as much as possible. Thus as shown in Figure 2.9, the ideal structure of a website would have much of the site's information accessible at the level (shown as horizontal bar). Structures that have multiple levels should concentrate the information at the first level when possible and at the level closest to the terminal nodes (at the bottom of the pyramid).

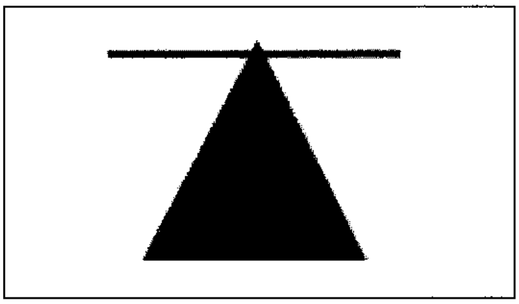


**Figure 2.9: The ideal web structure with multiple levels**

The arrangement of links can have a marked effect on search time and satisfaction. For example, it has been found that search time is significantly faster when links are grouped in columns rather than by rows (Nygren, 1996). However, as mentioned above, expandable link columns have been shown to decrease performance in terms of search time, errors, and number of clicks compared to bread crumbs, or simple link-column navigation (Maldonado & Resnick, 2002).

Moreover, as discussed in Usability news ( Bernard, 1999), experienced and novice users found specific links faster and were more satisfied with the structure of he site when the information is presented in columns according to their respective categories rather than when the links are presented in columns according to alphabetical listing of links. This is believed to occur because users have a difficult time trying to guess the appropriate link name in order to know where to initially look within he alphabetized column listing.

In the same study, users preferred to have all the menu links presented on one web page instead of initially showing only the link categories, which would then show the sub-category menus on mouse-over. The Author felt that the latter option of initially placing only the links categories would reduce "link crowding" on the screen to a more manageable number thus improving accuracy and satisfaction by making it easier to acquire the proper link. However, no significant differences between the former and the latter category options were found. Interestingly, a large portion of users sated that they would prefer the latter option if they were more familiar with the menu structure and the menu terms. Thus, one may want to have a full categorical link organization, but also have an option to initially show only the link categories, which would show the sub-menus on moue-over for frequent users of the site (Bernard, 1999).

In the specific placement of links within the website structure, (Kim and Yoo, 2000) found in a study of internet shopping mail sites that he combination to neighborhood links (links which move horizontally within the site). Top links (links which move the user upward to a predetermine destination, such as the homepage), an index links (links which go to the lowest level regardless of the current position, such as information on a specific product) significantly produce the greatest perception of navigation ease as well as general satisfaction. They also found that links which only moves up one level from its current position and down one level in a site causes a significantly lower perception of ease of navigation, as well as generally lowering the level of satisfaction with the site.

### 2.3.4 Web Graphics

Users have grown accustomed to looking in certain areas n a screen to find specific items (Bernard, 2001). Analyzing users'' expectation of where they expect specific web objects to be located revealed that generality:

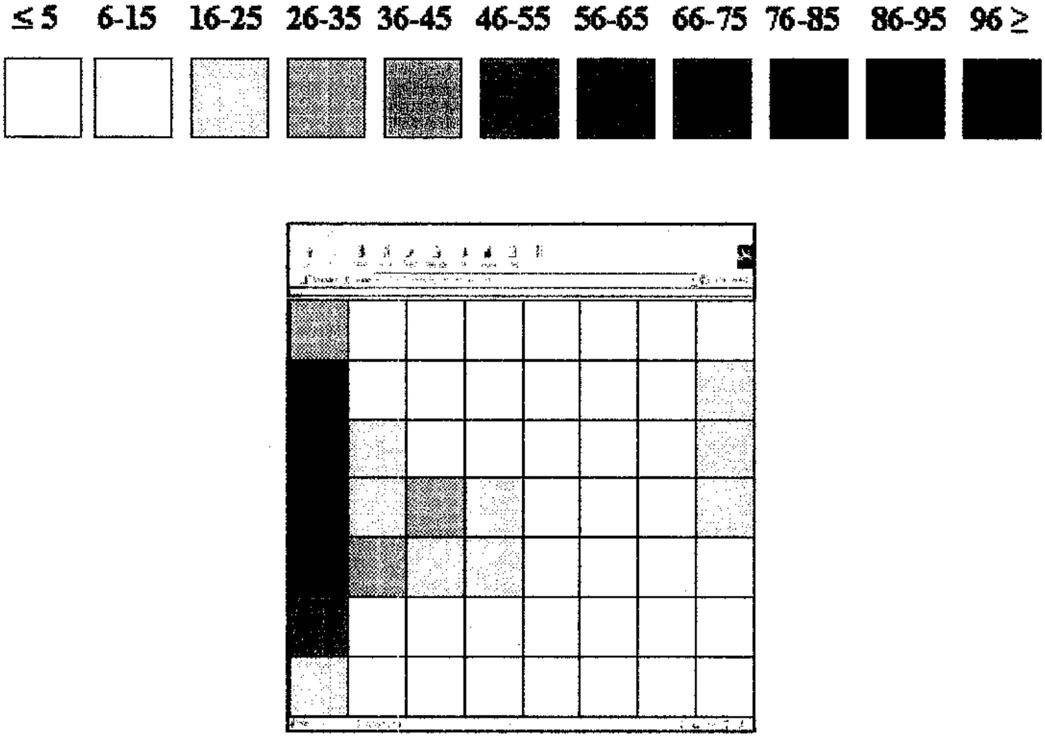


**Figure 2.10: Locations for internal web page links**

Internal web links expected to be located on the upper left side of the browser window (Figure 2.10).

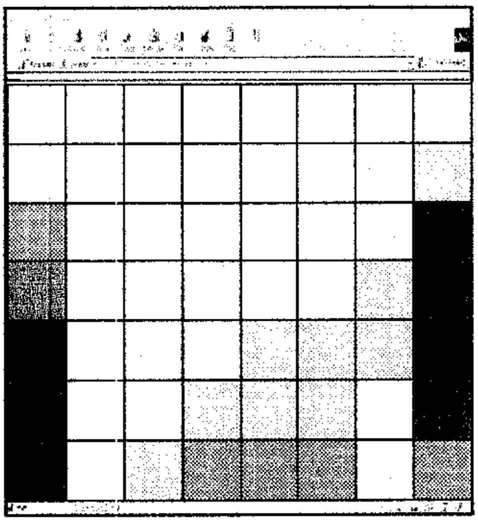


**Figure 2.11: Location for external website link**

External web links were expected to be located on the right side or lower left side of the browser window (Figure 2.11).

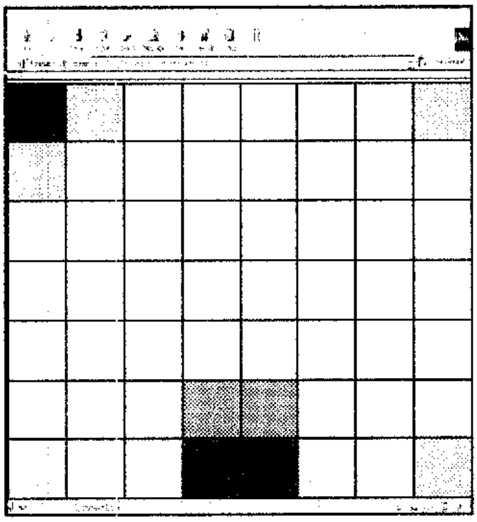


**Figure 2.12: Location for "back to home" link**

The "back to home" link was expected to be located at the top-left corner and the bottom-center of the browser window (Figure 2.12).

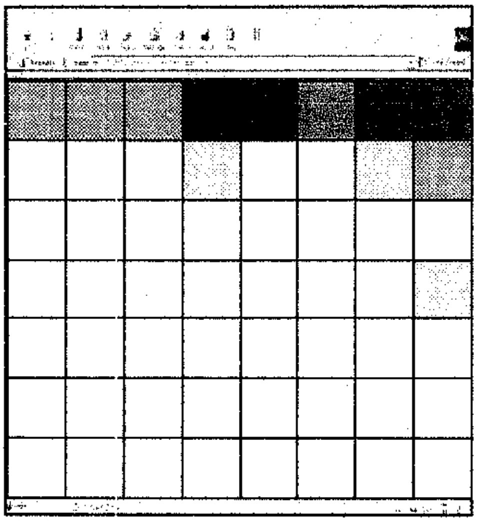


**Figure 2.13: Location for links to advertisement banners**

Links to advertisement banners were expected at upper center of a web page (Figure 2.13)

# CHAPTER 3
# METHODOLOGY AND PROJECT WORK

## 3. PROCEDURE IDENTIFICATION

### 3.1 Waterfall Model

For the project to be completed successfully, the method used also really important and draws many advantages. Even though there are many lifecycle models that can be chosen by the author to do the project, the author simply chooses the Waterfall model of system development that consists five basic phases which are planning, analysis, design and implementation.

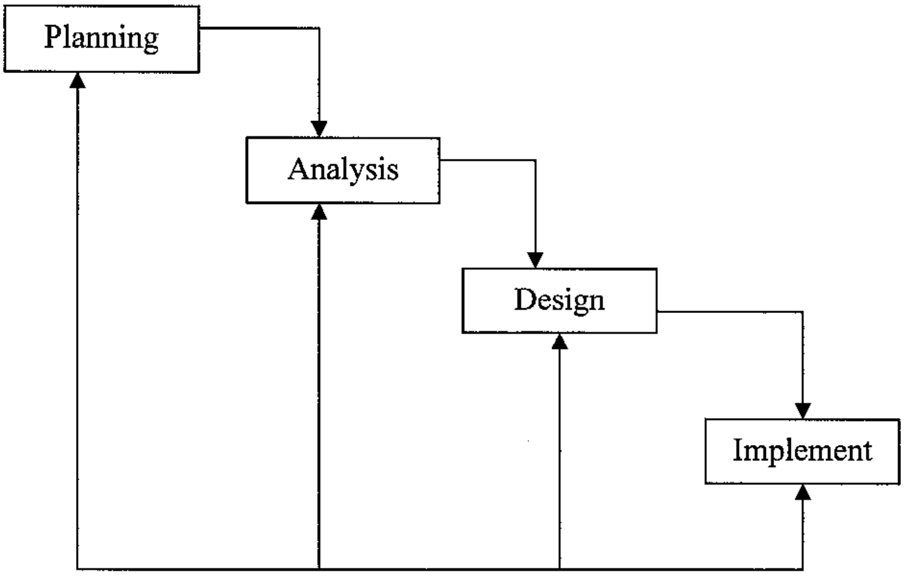


**Figure 3.1 Waterfall Model**

## 3.2 Significance of Waterfall Model

For this project the Waterfall model being used to perform all the activity, there would be several significant reasons that the model being chosen.

- Help minimize planning overhead
- Work well when quality requirement dominate cost and schedule
- The single requirements phase encourages specification of what the system is to do before deciding the system will do it (i.e. specification before design)
- The single design phase encourages planning of the system structure before building the components (i.e. specification before coding)
- The use of review at the end of each permits acquirer and user involvement
- The model permits early imposition of baseline and configuration control.

## 3.3 Waterfall Model Phases

### 3.3.1 Planning

In this phase, the author does the planning by proposed the project after make a project research to define the actual problem. After the title approved, the author continue doing the project scheduling to ensure that project can be completed within the time frame. The problem is analyzed in this phase before the recommended solution is given. Before that the author does see the supervisor to discuss about the planning activities in order for the author to proceed with the next phase of system development model.

### 3.3.2 Analysis

The Analysis Phase is also the part of the project where the identification of the overall direction that the project will take through the creation of the project strategy documents. Gathering requirements is the main attraction of the Analysis Phase. Depending on the complexity of the application, the process for gathering requirements has a clearly defined process of its own. This process consists of a group of repeatable processes that utilize certain techniques to capture, document, communicate, and manage requirements.

In this phase, an analysis of the problem statement and requirement analysis was done deeply. Many references that are relevant to the project have been gathered out for the literature review of the project. A critical review of the literature review been done in this phase. This is to identify the existing system structure and its limitation too. Besides that, this literature review will give a better and clear view of the entire related component that need to be considered in developing this system. The literatures reviewed included are textbooks, reference books, journals, white papers, web articles and Internet publication. The resources such as textbooks and reference books are borrowed from the library. Most of the journals and white papers published by companies and researches were mostly downloaded from the Internet. Articles from the Internets and other Internet publications were saved in HTML and Portable Document Format (PDF) form from the relevant sites. The findings are compiled and print onto papers and collected together into a folder as references through out the project development.

In this phase an elaboration and expanding of this system need to be done as it will illustrate the components and the requirement needs in this system.

### 3.3.3 Design

The third phase is one of important phase because; it is the stage where all the system specifications are translated into software representation. The software engineer at this stage is concerned with:

- Data structure
- Software architecture
- Algorithmic detail
- Interface representations

The Design Phase is where you look at the many potential solutions and narrow down the choices to determine the most effective and efficient way to construct the solution.

In this phase, the system flow and design process need to be completed as it will contribute to the success of the next phase which is the construction phase. The step by step process flow of the system is a very crucial design to be done as it show the process behind the working system that will be develop in this project.

### 3.3.4 Implementation

In this phase the designs is translated into software domain. Testing is also called quality assurance (QA), it includes not only unit tests, but also integration test that exercise the subsystem.

## 3.4 Development Tools

### 3.4.1 Software

- Microsoft Visual C++.NET

Microsoft Visual C++ uses an integrated development environment known as Visual Studio. Visual Studio is a development environment that can be used with a suite of programming languages. Visual Studio provide a seamless environment that can be used to maintain source code, design a user interface, and compile and link projects, as well as for debugging and testing. The development environment that this software has is known as an integrated development environment, or IDE, since all the tools that are needed are integrated into one application.

- Microsoft Word

Microsoft Word is one of word authoring tools. In this project, Microsoft Word is being used to write all related documentation such as to write progress report, final report and weekly report.

### 3.4.2. Hardware

The operating system that will be used for the project is Windows XP Home Edition. The standard Multimedia personal Computer with capabilities of 15 inch (1024 x 768 resolutions) monitor, 240 MB Random Access Memory (RAM), 40 GB Hard Disk and Intel Pentium Centrino 1400 MHz processor will be used as the workstation for the development.

# CHAPTER 4
# RESULT AND DISCUSSION

### 4 Findings and Discussion

The Web has created an explosion of freely available rich (and not so rich) information. On top of this, even the most basic computer has gigabytes of storage space on which important information is stored and can be hard to find. Yet despite all this available data, there is a problem: How do you get the exact information you want, in the format you want, quickly, easily and without great expense? Solving this problem is critical to staying competitive.

### 4. 1 Without extraction tools

Tools are needed to manage all available information including the Web, subscription services, and internal data stores. Without an extraction tool (a product specifically designed to find, organize, and output the data you want), you have very poor choices for getting information. The user choices are:

- Use search engines

Search engines help find some Web information, but they do not pinpoint information, cannot fill out web forms they encounter to get you the information you need, are perpetually behind in indexing content, and at best, can only go two or three levels deep into a Web site. And they cannot search file directories on your network.

- Manually surf the Web and file directories

The need of worker effort to give their ability in searching desired data, the process of finding desired information would be tedious, costly, error prone, and very time consuming. Web users have to read the content of each page to see if it suits their needs, whereas a computer is simply matching patterns, which is more effective.

- Create custom programming

Custom programming is costly, can be buggy, requires maintenance, and takes time to develop. Plus the programs must be constantly updated as the location of information frequently changes.

Inefficient methods means the information analyst spends time finding, collecting, and aggregating data instead of analyzing data and gaining the competitive edge. This also affects the application programmer who has to spend time developing extraction tools instead of developing tools for the core business.

### 4.2 New solutions improve productivity

Extraction tools using a concise notation to define precise navigation and extraction rules greatly reduce the time spent on systematic collection efforts. Tools that support a variety of format options provide a single development platform for all collection needs regardless of electronic information source. Early attempts at software tools for "Web harvesting" and unstructured data mining emerged, and started to get the attention of information professionals. These products did a reasonable job of finding and extracting Web information for intelligence gathering purposes. But this was not enough. Organizations needed to reach the "deep Web" and other electronic information sources, capabilities beyond simplistic Web content clipping. A new generation of information extraction tools is markedly improving productivity for information analysts and application developers.

### 4.3 Uses for extraction tools

The most popular applications for information extraction tools remain competitive intelligence gathering and market research, but there are some new applications emerging as organizations learn how to better use the functionality in the new generation of tools.

- **Deep Web price gathering**

The explosion of e-tailing, e-business, and e-government makes a plethora of competitive pricing information available on Web sites and government information portals. Unfortunately, price lists are difficult to extract without selecting product categories or filling out Web forms. Also, some prices are buried deep in .pdf documents. Automated forms completion and automated downloading are necessary features to retrieve prices from the deep Web.

- **Primary research**

Message boards, e-pinion sites, and other Web forums provide a wealth of public opinion and user experience information on consumer products, air travel, test drives, experimental drugs, etc. While much of this information can be found with a search engine, features like simultaneous board crawling, selective content extraction, task scheduling, and custom output reformatting are only available with extraction tools.

- **Content aggregation for information portals**

Content is exploding and available from Web and non-Web sources. Extraction tools can crawl the Web, internal information sources, and subscription services to automatically populate portals with pertinent content such as competitive information, news, and financial data.

- **Scientific research**

Scientific information on a given topic (such as a gene sequence) is available on multiple Web sites and subscription services. An effective extraction tool can automate the location and extraction of this information and aggregate it into a single presentation format or portal. This saves scientific researchers countless hours of searching, reading, copying, and pasting.

- **Business activity monitoring**

Extraction tools can continuously monitor dynamically changing information sources to provide real time alerts and to populate information portals and dashboards.

### 4.4 Extraction tools versus search engines

What is the difference between an information extraction tool and a search engine? The simple answer is that extraction tools pick up where search engines leave off, doing the work the search engine is not capable of. Table 1 displays a side-by-side comparison:

| **Information gathering mechanism** | **Locate** | **Pinpoint** | **Extract** | **Integrate** |
|---|---|---|---|---|
| Extraction tool | **automated** | **automated** | **automated** | **automated** |
| Search engine | **automated** | manual | manual | manual |

**Table 1: Extraction tools versus Search engine**

#### 4.4.1 Search engines

Search engines locate information and point to it. They typically go no deeper than two or three levels into a Web site to find information and then return URLs, meta descriptions, and meta keywords. The meta descriptions and keywords can be bogus, because some webmasters load their meta data with popular descriptions and keywords in order to create hits, and search engines cannot distinguish the difference. The search engine cannot do anything beyond the simple matching of keywords to return URLs. If you are using a search engine for data gathering, tasks such as pinpointing, extraction, and integration of useful information, have to be accomplished by a person or people who complete the following steps:

- Skim the content until the information is found
- Mark the information (usually with a mouse)
- Copy information
- Switch to another application (such as a spreadsheet or database)
- Paste the information into that application

Taking it a step further, many Web sites or subscription services require one or more manual entries prior to retrieving and displaying information, further complicating and elongating the process. Automated forms discovery and completion is something most search engines cannot do, and add another task for the researcher.

### 4.4.2 Extraction tools

Extraction tools automate the full process of gathering, pinpointing, and outputting data, thereby freeing resources from these tasks. A robust extraction tool should be able to perform all the tasks you need, quickly and easily.

**Attributes of a comprehensive extraction tool**

Not all information extraction tools are alike. You should consider the following when evaluating an extraction product.

- **Precision and automation**

  There are three types of methods you can use to locate and extract data. Some extraction tools use Artificial Intelligence (AI) techniques. While interesting and highly automated, these tools are often imprecise. Some tools use a drag-and-drop GUIs, which provides precision, but are not automatic and can be impractical if the data you are seeking resides on a variable number of pages with inconsistent formats. The most effective tools use a concise notation capable of calling on a variety of precise extraction techniques depending on the structure of the information solving both the automation and the precision problems.

- **Integration with Application Programming Interfaces (APIs)**

  Application programmers or Web masters may want to incorporate content into their applications. To enable this, extraction tools should provide robust API's for the popular computing environments (ActiveX, COM, Java, C++, VBA, SOAP).

- **Integrated navigation and extraction**

An extraction tool should have integrated navigation and extraction. A superior extraction tool uses a seamless method to both navigate and extract data equally well. An inferior extraction tool may be effective once it gets to a page, but cannot navigate well to the page in the first place. This is the issue with tools using both AI and drag-and-drop GUIs. To deal with this deficiency, these tools "bolt on" scripting languages to handle navigation. The result of a mix of technologies for navigation and extraction is an application that is clunky, slow, and difficult to maintain, particularly if the navigation is page-content sensitive.

- **Scalability**

Information extraction applications can be very resource and time intensive, therefore extraction tools must be scalable and distributable. A superior tool can be distributed across multiple processors, does not limit the number of processors per application, and can process multiple applications in parallel.

- **Identity protection**

For a variety of reasons, organizations may not want to be identified when visiting Web sites. Depending on the application and traffic volume, extraction tools need to provide options for identity protection. For high performance and ease of implementation, this capability should be tightly integrated.

- **Multi-format support**

Information comes from a variety of places and in a variety of formats. An effective tool should extract data from all the commonly used file formats (e.g., .pdf, .doc, .xls, .html). It should be able to find them both on the Web as well as from sources outside the Web (internal data stores and subscription services). Lastly it should be able to directly integrate the output into popular presentation or storage formats.

### 4.5 Document Object Model (DOM)

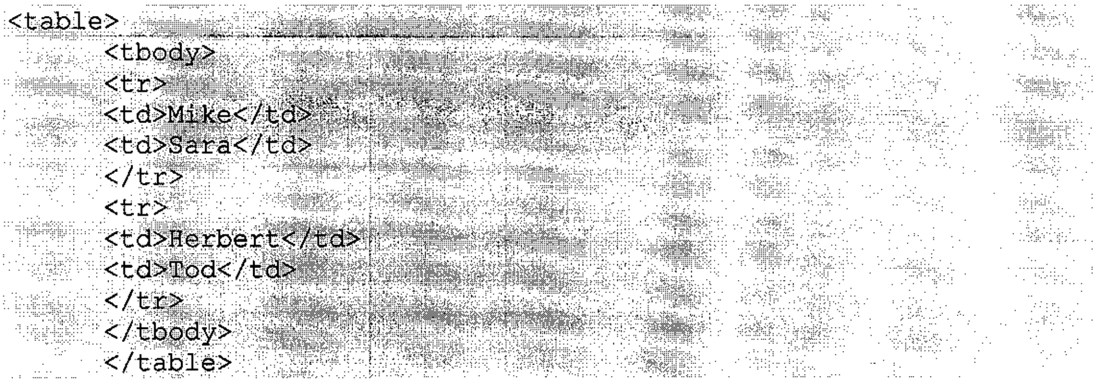
There were several methods or techniques that being used in order to extract web content which being used by previous researcher but the obvious and usual model that being used is using Document Object Model.

Document Object Model (DOM) is an application programming interface (API) for valid HTML and well-formed XML documents. It defines the logical structure of documents and the way a document is accessed and manipulated. In the DOM specification, the term "document" is used in the broad sense - increasingly, XML is being used as a way of representing many different kinds of information that may be stored in diverse systems, and much of this would traditionally be seen as data rather than as documents. Nevertheless, XML presents this data as documents, and the DOM may be used to manage this data.

With the Document Object Model, programmers can build documents, navigate their structure, and add, modify, or delete elements and content. Anything found in an HTML or XML document can be accessed, changed, deleted, or added using the Document Object Model, with a few exceptions - in particular, the DOM interfaces for the XML internal and external subsets have not yet been specified.

As a W3C specification, one important objective for the Document Object Model is to provide a standard programming interface that can be used in a wide variety of environments and applications. The DOM is designed to be used with any programming language.

The DOM is a programming API for documents. It is based on an object structure that closely resembles the structure of the documents it models. For instance, consider this table, taken from an HTML document:



```
<table>
          <tbody>
          <tr>
          <td>Mike</td>
          <td>Sara</td>
          </tr>
          <tr>
          <td>Herbert</td>
          <td>Tod</td>
          </tr>
          </tbody>
          </table>
```

A graphical representation of the DOM of the example table is shown on figure 4.1 below:

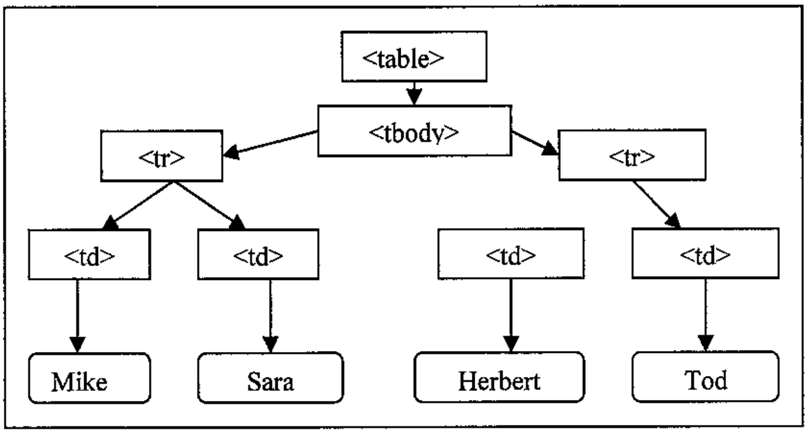


**Figure 4.1**

In the DOM, documents have a logical structure which is very much like a tree; to be more precise, which is like a "forest" or "grove", which can contain more than one tree. Each document contains zero or one doctype nodes, one root element node, and zero or more comments or processing instructions; the root element serves as the root of the element tree for the document. However, the DOM does not specify that documents must be implemented as a tree or a grove, nor does it specify how the relationships among objects be implemented. The DOM is a logical model that may be implemented in any convenient manner. In this specification,the term “structure model” is being used to describe the tree-like representation of a document. The term "tree" also being used when referring to the arrangement of those information items which can be reached by using "tree-walking" methods; (this does not include attributes). One important property of DOM structure models is structural isomorphism: if any two Document Object Model implementations are used to create a representation of the same document, they will create the same structure model, in accordance with the XML Information Set (Info Set).

There may be some variations depending on the parser being used to build the DOM. For instance, the DOM may not contain whitespaces in element content if the parser discards them.

The name "Document Object Model" was chosen because it is an "objec model" in the traditional object oriented design sense: documents are modeled using objects, and the model encompasses not only the structure of a document, but also the behavior of a document and the objects of which it is composed. In other words, the nodes in the above diagram do not represent a data structure, they represent objects, which have functions and identity. As an object model, the DOM identifies:

- the interfaces and objects used to represent and manipulate a document
- the semantics of these interfaces and objects - including both behavior and attributes
- the relationships and collaborations among these interfaces and objects

The structure of SGML documents has traditionally been represented by an abstract data model, not by an object model. In an abstract  data model, the model is centered around the data. In object oriented programming languages, the data itself is encapsulated in objects that hide the data, protecting it from direct external manipulation. The functions associated with these objects determine how the objects may be manipulated, and they are part of the object model.

## 4.6 Product Result

The extractor is being built using Visual C++.Net and the interface is being shown as below.

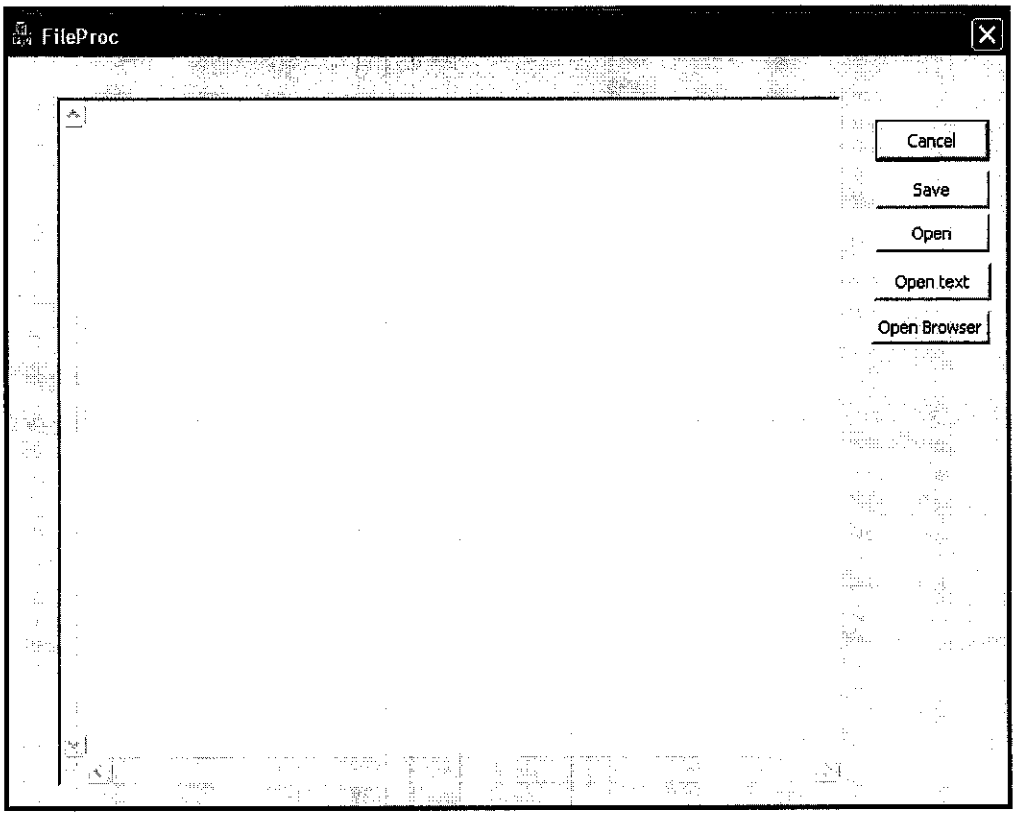
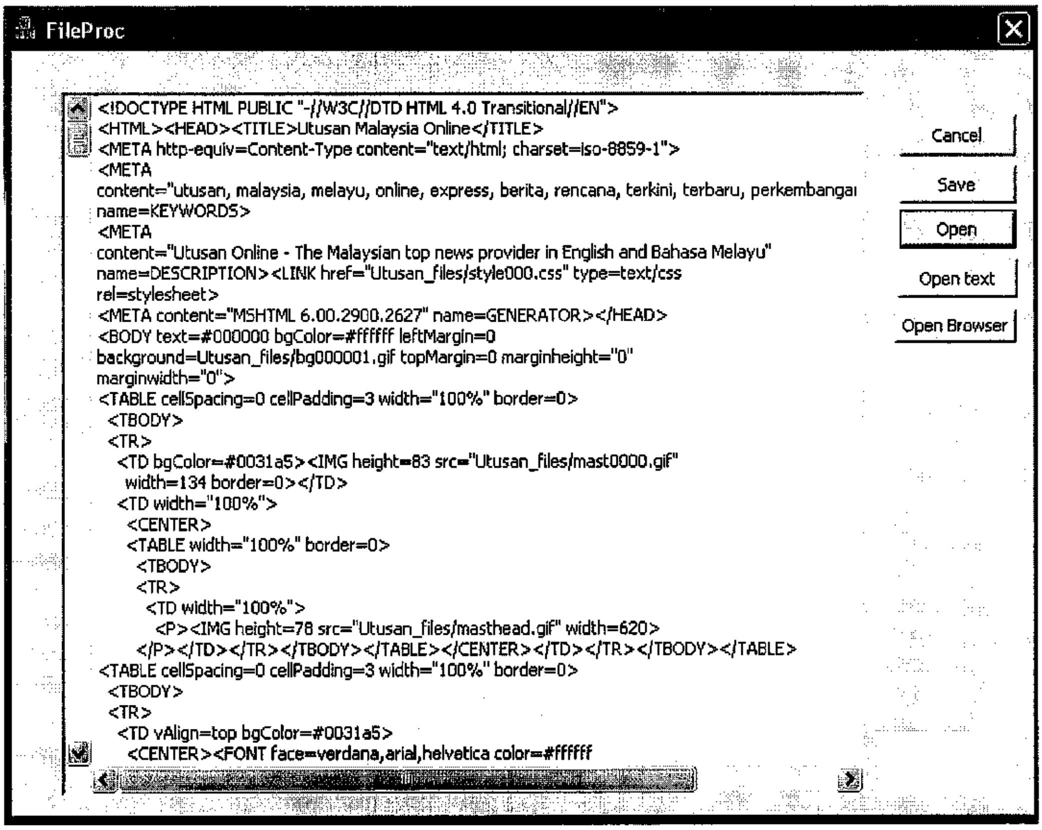


Figure 4.2

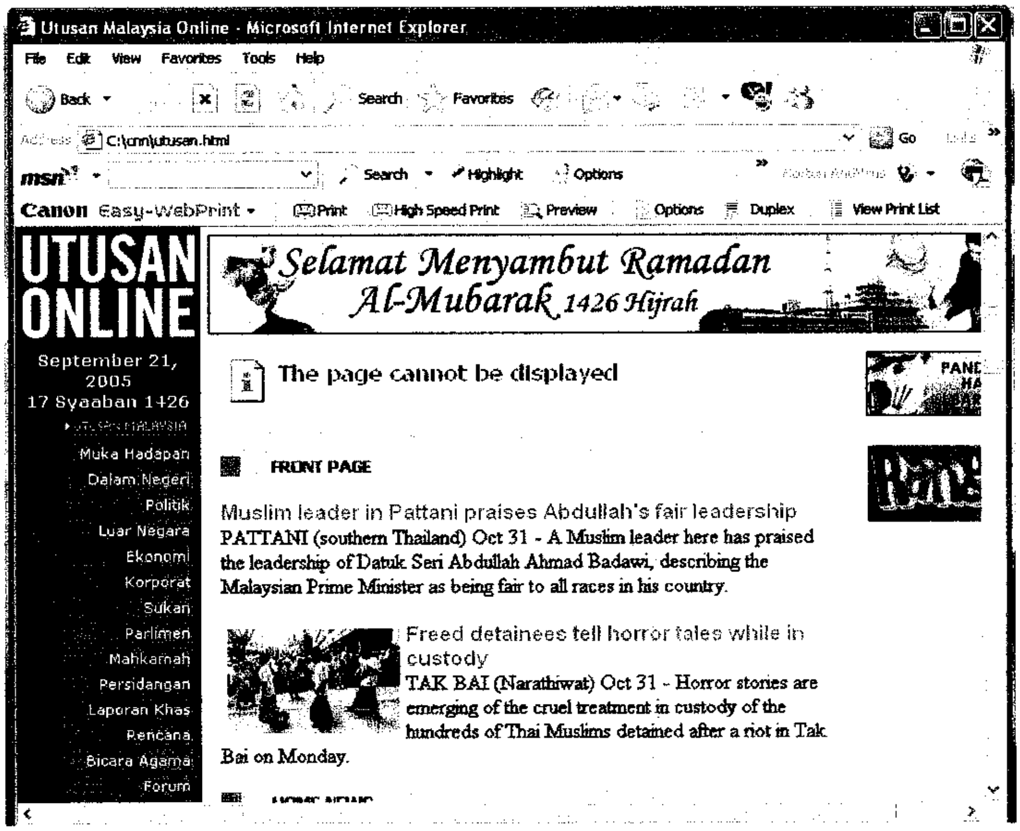
On figure 4.2, there would be a display area where the user can click the button and the html code will be displayed. The code is come from the news page that has being selected.



```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<HTML><HEAD><TITLE>Utusan Malaysia Online</TITLE>
<META http-equiv=Content-Type content="text/html; charset=iso-8859-1">
<META
content="utusan, malaysia, melayu, online, express, berita, rencana, terkini, terbaru, perkembangai
name=KEYWORDS>
<META
content="Utusan Online - The Malaysian top news provider in English and Bahasa Melayu"
name=DESCRIPTION><LINK href="Utusan_files/style000.css" type=text/css
rel=stylesheet>
<META content="MSHTML 6.00.2900.2627" name=GENERATOR></HEAD>
<BODY text=#000000 bgColor=#ffffff leftMargin=0
background=Utusan_files/bg000001.gif topMargin=0 marginheight="0"
marginwidth="0">
<TABLE cellSpacing=0 cellPadding=3 width="100%" border=0>
  <TBODY>
  <TR>
    <TD bgColor=#0031a5><IMG height=83 src="Utusan_files/mast0000.gif"
      width=134 border=0></TD>
    <TD width="100%">
      <CENTER>
      <TABLE width="100%" border=0>
        <TBODY>
        <TR>
          <TD width="100%">
            <P><IMG height=78 src="Utusan_files/masthead.gif" width=620>
        </P></TD></TR></TBODY></TABLE></CENTER></TD></TR></TBODY></TABLE>
<TABLE cellSpacing=0 cellPadding=3 width="100%" border=0>
  <TBODY>
  <TR>
    <TD vAlign=top bgColor=#0031a5>
      <CENTER><FONT face=verdana,arial,helvetica color=#ffffff
```

**Figure 4.3**

On figure 4.3, after the button open being clicked the initial code for the selected web pages is being displayed.



UTUSAN ONLINE

September 21, 2005
17 Syaaban 1426

Muka Hadapan
Dalam Negeri
Politik
Luar Negara
Ekonomi
Korporat
Sukan
Parlimen
Mahkamah
Persidangan
Laporan Khas
Rencana
Bicara Agama
Forum

Selamat Menyambut Ramadan Al-Mubarak 1426 Hijrah

The page cannot be displayed

FRONT PAGE

Muslim leader in Pattani praises Abdullah's fair leadership

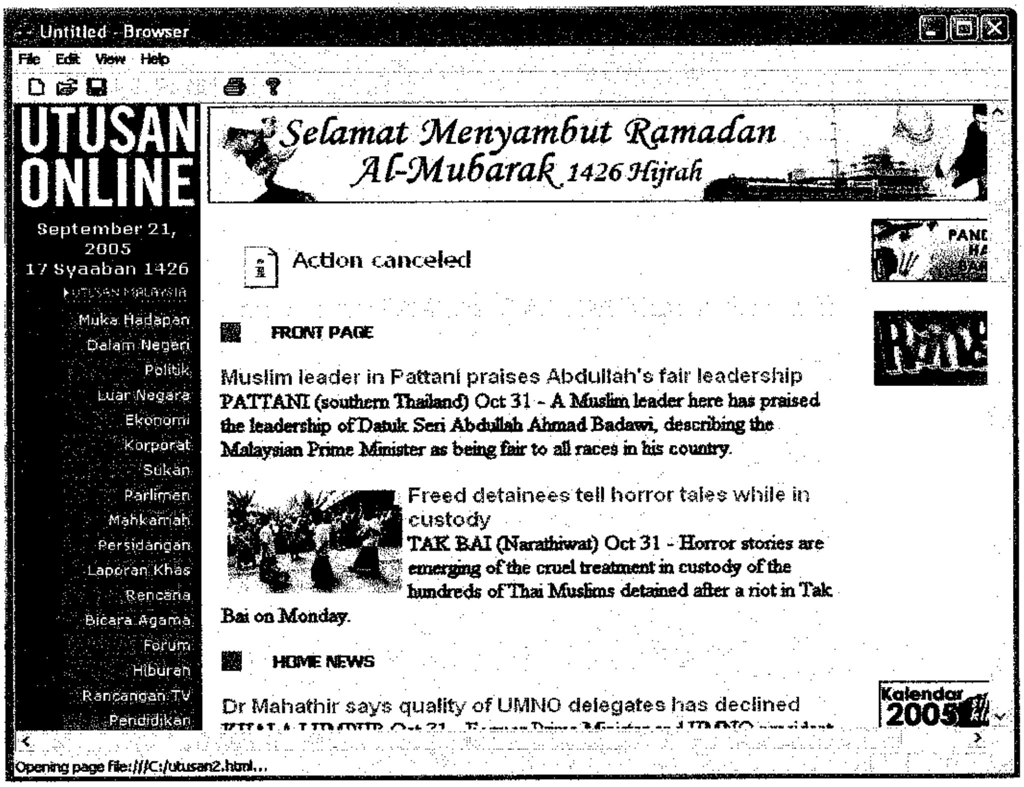
PATTANI (southern Thailand) Oct 31 - A Muslim leader here has praised the leadership of Datuk Seri Abdullah Ahmad Badawi, describing the Malaysian Prime Minister as being fair to all races in his country.

Freed detainees tell horror tales while in custody

TAK BAI (Narathiwat) Oct 31 - Horror stories are emerging of the cruel treatment in custody of the hundreds of Thai Muslims detained after a riot in Tak Bai on Monday.

**Figure 4.4**

On figure 4.4, the news web page is in the original form



Action canceled

FRONT PAGE

Muslim leader in Pattani praises Abdullah's fair leadership

PATTANI (southern Thailand) Oct 31 - A Muslim leader here has praised the leadership of Datuk Seri Abdullah Ahmad Badawi, describing the Malaysian Prime Minister as being fair to all races in his country.

Freed detainees tell horror tales while in custody

TAK BAI (Narathiwat) Oct 31 - Horror stories are emerging of the cruel treatment in custody of the hundreds of Thai Muslims detained after a riot in Tak Bai on Monday.

HOME NEWS

Dr Mahathir says quality of UMNO delegates has declined

Opening page file:///C:/utusan2.html...

**Figure 4.4**

On figure 4.4, the content of the news web page is being highlighted

# CHAPTER 5

# CONCLUSION AND RECOMMENDATION

## 5 Relevancies to the Objective

The idea of this application is to provide an application that can analyze news domain from Internet web pages. Several news web pages will be collected which may consist of many type of web structure. The process of collecting the web pages will be done. The process of developing the application will involve of analyzing and determining what are the related data, which related to the news specifically. The purpose of this project is to ease the web user in order to retrieve the information that they wanted. The domain web pages can be any domain, but for this project news domain will be chosen. This project hopefully will help the web user to gain their desired information faster and more accurately.

## 5.1 Recommendation

The Internet can contain many of information and having many types of users. The intention for web navigation maybe various depending on the users need. The process of navigating and retrieving the information on the Internet sometimes can take many times and the content extraction may help in the process.

The process of highlighting the relevance information is one of way that the author found in the process of web content extraction. For future enhancement the process of extracting the content may be can be represented in more precise, where the initial content can be extracted and the data can be presented in a single form and the process can be fasten. The user can easily read out all the related information that they desired as the information being organize more neatly and nicely.

# REFERENCES

Atzeni, P., Mecca, G. and Merialdo, P., *To Weave the Web - In Proceedings of the 23rd International Conference on Very Large Databases* (VLDB'97), 1997

Aidan Finn, Nicholas Kushmerick and Barry Smyth. *Fact or fiction: Content classification for digital libraries.* In Joint DELOS-NSF Workshop on Personalisation and Recommender Systems in Digital Libraries (Dublin), 2001.

Baeza-Yates,R. and Ribeiro-Netro.B, *Modern information Retrievel.* Addison-Wesley,Harlow, England, 1st edition,1999.

Chen, Y., Ma, W.Y., and Zhang, H.J. "*Detecting Web Page Structure for Adaptive Viewing on Small Form Factor Devices*". Proc. WWW'03 Budapest, Hungary, May 2003.

Florecu,D., Levy,A. , and Mendelzon,A.. *Database techniques for the world-wide web: a survey.* SIGMOD Rec., 27(3):59-74,1998.

Jacob Nielson, *Failure of Corporate Websites.*,1998.

Kaasinen,E., Aaltonen,M., Kolari,J.,Melakoski,S. and T. Laakko,T. "*Two Approaches to Bringing Internet Services to WAP Devices*". In Proc. of 9th Int. World-Wide Web Conf., 2000.

Laender,A., Ribeiro-Neto,A,B., Silva, and J.S. Teixeira,J.S. *A brief Survey of Web data extraction tools.*SIGMOD Record, 31(2):84-93,2002.

Micheal L Bernard,*Criteria for optimal web design*, 2003.

Rahman, A. F. R., Alam,H. and Hartono,R. "*Content Extraction from HTML Documents". In 1st Int. Workshop on Web Document Analysis* (WDA2001), 2001.

W3C-Semantic Web, http://www.w3.org/2001/sw/ Proceedings of the 23rd International Conference on Very Large Databases (VLDB'97), 1997

http://www.w3.org/DOM/