

Online News Headline Extraction

by

Anis Akmal Binti Anuar

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

MAY 2011

Universiti Teknologi PETRONAS
Bandar Seri Iskandar
31750 Tronoh
Perak Darul Ridzuan

CERTIFICATION OF APPROVAL

Online News Headline Extraction

by

Anis Akmal Binti Anuar

A project dissertation submitted to the
Computer Information Science Department
Universiti Teknologi PETRONAS
in partial fulfilment of the requirement for the
BACHELOR OF TECHNOLOGY (Hons)
(Information and Communication Technology)

Approved by,



(Ms. Amy Foong Oi Mean)

UNIVERSITI TEKNOLOGI PETRONAS
TRONOH, PERAK
May 2011

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 here in have not been undertaken or done by unspecified sources or persons.



ANIS AKMAL BINTI ANUAR

ABSTRACT

This paper presents the online headline news extraction application. According to research, today's online news has grown 11 % year over year. Users nowadays are overwhelmed with too much on the internet. The current online news also is not visible for user to read the news; this is because it is full of the advertisement and other unrelated thing besides the news itself. This paper purposes the proposal of an E-Headlines News Extraction Framework that illustrated the extracted information on the news. This project will only cover the news reported or news available on the local online English newspaper and at the mean time try to extract the headlines of the news first. At the end of the project, it will highlight the application that can illustrate the extracted information on the news.

ACKNOWLEDGMENTS

This work has benefited from the input of many people. The author would like to thank Ms Amy Foong Oi Mean for her guide and contribution throughout the project completion and her valuable feedback on the project. Special thanks to Ms Siti Rohkmah also for her guidance throughout the project. Particular thanks should go to author's friends, Norfarahin Adiba Abd Kadir, Nik Nor Ernina binti Nik Ab Rahman, Wan Aathena Wan Ahmad Marzuki, Siti Ain Nurena binti Mohd Nasir, Nurul Ain binti Md Yazid, and Nur Hidayah binti Fadzil for their support and motivation towards the completion of this project.

TABLE OF CONTENTS

ABSTRACT	i
ACKNOWLEDGMENT	ii
CHAPTER 1: INTRODUCTION	1
1.1 Background	1
1.2 Problem Statement	4
1.3 Objective	4
1.4 Scope of Study	5
CHAPTER 2: LITERATURE REVIEW	6
CHAPTER 3: METHODOLOGY / PROJECT WORK	12
3.1 Project Methodology	12
3.1.1 Planning	14
3.1.2 Analysis	14
3.1.3 Design	15
3.1.4 Implementation	15
3.1.5 System Testing	16
3.1.6 Final System	16
3.2 Proposed Framework	17
CHAPTER 4: RESULT AND DISCUSSION	19
4.1 System Architecture	19
4.2 System Process	20
CHAPTER 5: CONCLUSION AND RECOMMENDATION	31
REFERENCES	32
APPENDICES	35

LIST OF FIGURES

Figure 1.1	WordPress template hierarchy	2
Figure 2.1	The architecture of NEXUS	8
Figure 2.2	Real-time event extraction processing chain	9
Figure 2.3	Google news	11
Figure 3.1	System methodology; prototyping-based	13
Figure 3.2	Proposed framework	17
Figure 4.1	System Architecture	19
Figure 4.2.1	cPanel	20
Figure 4.2.2	cPanel Homepage	21
Figure 4.2.3.1	MySQL Database	22
Figure 4.2.3.2	Current Database	22
Figure 4.2.3.3	phpMyAdmin Database	23
Figure 4.2.4.1	New Straits Times online	24
Figure 4.2.4.2	Top news on New Strait Times	24
Figure 4.2.4.3	The Star online	25
Figure 4.2.4.4	Top news on The Star	25
Figure 4.2.4.5	Malay Mail online	26
Figure 4.2.4.6	Top news on Malay Mail	26
Figure 4.2.5.1	Aggregator widget	27
Figure 4.2.5.2	Home page feeds	28
Figure 4.2.6.1	About Xtractly News	29
Figure 4.2.6.2	Snapshot of the interface	30

CHAPTER 1

INTRODUCTION

1.1 Background

With the outburst of the World Wide Web, a wealth of data on almost every subject has become available online. According to the statistical results by Miniwatts Marketing Group, the growth of web users during this decade is over 200% and there are more than 1 billion Internet users from over 233 countries and world regions [1]. Generally, users retrieve Web data by browsing and keyword searching. Though all searches will produce links, there are limitations, and disadvantages in methodology. Data on the internet are not structured or ordered as from databases. Gathering and formatting data in desired way is what data extraction is all about. Data extraction is the ability to retrieve data from web, and to transform and transfer it in a pre-determined way.

Cunningham (2004) said that “Information Extraction (IE) is a technology based on analyzing natural language in order to extract snippets of information“ [2] . IE automatically finds and extracts the relevant information to help people overcome information overloading and it is usually used to extract specific information that people are interested in. According to Pinheiro et al, “IE system requires a module for the semantic analysis to understand and extract the required information” [3].

According to Man I, Zhiguo and Maybin (2008), the information source can be classified into three main types, including free text, structured text and semi-structured text [4] . Originally, the extraction system focuses on free text extraction. Natural Language Processing (NLP) techniques are developed to extract this type of information, The structured information usually comes from databases, which provide rigid or well defined formats of information, therefore, The other type is the semi structured information, which falls between free text and structured information. Web pages are a typical example of semi-structured information.

This system is using WordPress. WordPress is an open source Content Management System (CMS), often used as a blog publishing application, powered by PHP and MySQL. It has many features including a plug-in architecture and a template system. WordPress is used by over 13% of the 1,000,000 biggest websites.

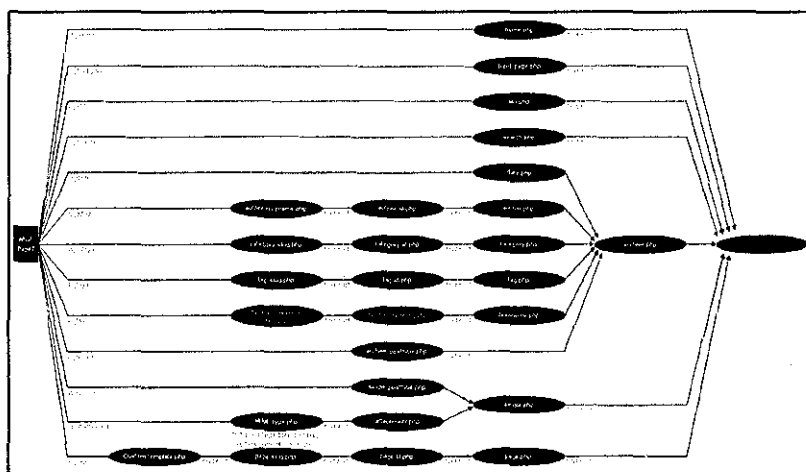


Figure 1.1 : WordPress template hierarchy

What is news aggregator? In general Internet terms, a news aggregation website is a website where headlines are collected, usually manually, by the website owner. In computing, a feed aggregator, also known as a feed reader, news reader, RSS reader or simply aggregator, is client software or a Web application which aggregates syndicated web content such as news headlines, blogs, podcasts, and blogs in a single location for easy viewing.

Aggregator is a built in tools in the WordPress theme. The main function of aggregator is to subscribe the feed from others website. Aggregators reduce the time and effort needed to regularly check websites for updates, creating a unique information space or “personal newspaper”. Once subscribed to a feed, an aggregator is able to check for new content at user-determined intervals and retrieve the update.

Really Simple Syndication (RSS) is a family of web feed formats used to publish frequently updated works- such as blog entry, news headline, audio and video in a standardizes format. An RSS document which is called a feed, web feed or channel includes full of summarized text, plus metadata such as publishing dates and authorship. Web feed benefits publishers by letting them syndicate content automatically. They benefits user who want to subscribe to timely updates from favored websites or to aggregate feeds from many site to one place.

This project will propose a model or framework of web information extraction unit for extracting headline on the online news and developed a simple prototype system that can illustrate the extracted headline on the online news.

1.2 Problem Statement

According to research group by Nilsen/Net ratings a global leader in the internet media and market research, today's online news have grew 11 % year over year [5] . Internet articles on news keep increasing. Users nowadays are overwhelmed with too much on the internet. The current online news is not visible for user to read the news; this is because it is full of the advertisement and other unrelated thing besides the news itself.

People are getting hectic day by day, so they need something simple that can help them getting update of the current issues without spending much time on it. This project will come out with application that can extract the headlines and represent it in more simple way to make it easy to be captured by the user. By having this, the user can just review the headlines of the news for the day and choose which one that suit their interest and read it.

1.3 Objective

The objective of this system is to propose an E-Headline News Extraction Framework that illustrates the extracted information on the news.

1.4 Scope of Study

To cope with the time and cost given, this project will only cover the news reported or news available on the local online English newspaper. At the mean time this project will first extract the news headlines and display it , as it was easy for the user to review the headlines and just view the news that they are intrested in.

CHAPTER 2

LITERATURE REVIEW

Information extraction is the process of identifying a pre-specified set of key data elements from a free-text data source [6,7,8]. In the survey perform by Norshuhanani (2009) for the counter- terrorism show the proposed information extraction framework consist of various technique text pre-processing, name entity recognition, conference resolution, entity extraction and linked list [9] .

- **Text Pre-processing**

The preprocessing algorithm should consider the cleansing of the text, only text-based (no images, unwanted whitespace), text tokenization and part-of-speech tagging.

- **Name Entity Recognition (NER)**

NER classified entities in the text into predefined categories such as names of person, organizations, locations, dates, time, quantities, monetary values, percentages and etc.

- **Conference Resolution**

Conference resolution involves identifying different description of the same entity that have been identified by the NER module in the different parts of a text. It requires deeper analysis on syntactic and semantic chains of the information.

- **Entity Extraction**

Entity extraction algorithm identifies relevant facts in texts and their classification into a set of predefined categories of interest.

- **Linked List**

Link analysis is a technique of the data mining field concerns with extracting useful information from a large dataset of association between entities.

In the other hand, the field study carry out, Robert Dale et al (2003) found the new approach to information extraction that neatly integrates top-down hypothesis driven information with bottom up data driven information called path-merging [10]. The template will provides top-down hypotheses as to the information find in text, the name entities identified in the text provide bottom up data that is merged with the hypotheses.

However Bettina, Sergio and Andrew (2009), said that the effective solution to automate information extraction from web pages is represented by wrappers [11]. Wrapper associates a web page with an XML document. Bettina et al proposed schema-based wrapping approach that can range from capability of simply guiding and scheming the extraction and integration of required data from HTML document to the specification of structured yet easy extraction rule.

Conventional method for data extraction can be generally divided into four categories.; information extraction based on natural language processing, information extraction based on the wrapper summing up the rule, information extraction based on ontology and information based on HTML structure. Huan and Yang (2010) proposed the simple tree matching algorithm to extract data record from similar Web page, it will integrates the structural similarity of web page and its correlation with Xpath [12] . The drawback of this method is the algorithm can only extract the data from web with the same structured only.

Hristo Tanev et al (2008) have come out with real time news event extraction for extracting the violent and disaster events from online news without using so much linguistic sophistication accurately and efficiently [13]. The news article have been collected through the Internet with the Europe Media Monitor (EMM). EMM is a web based news aggregation system that collects 40000 news articles from 1400 news sources in 35 languages each day. And then the input-data is geo-tagged, categorized and grouped into news cluster. Next, each cluster is processed by NEXUS (News cluster Event eXtraction Using language Structures). Their core slots are: date and location, number of killed and injured, kidnapped people and type of event.

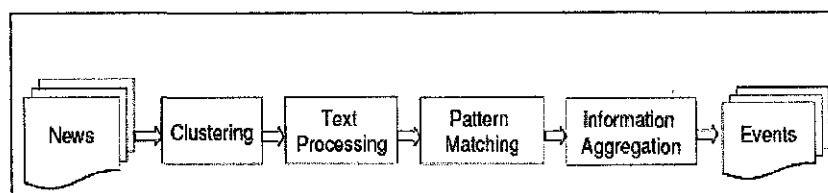


Figure 2.1: The architecture of NEXUS

On the other hand, the research conducted by Jakub Piskorski et al (2008), found a new approach to present a real-time and multilingual news event extraction system that is a real-time event extraction processing chain [14]. It is capable of competently processing news in English, Italian, French and Arabic, which present descriptions of the latest crisis-related events around the world with a 10-minute delay. The results of the live event extraction are presented via a publicly reachable web page and can be also accessed with the Google Earth application.

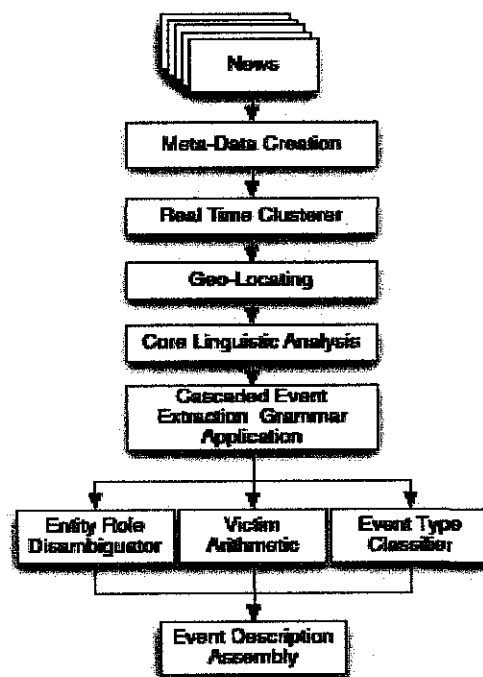


Figure 2.2: Real-time event extraction processing chain

Shuyi Zeng et al (2007) proposed a template-independent news extraction approach based on visual consistency [15]. They represent a page as a visual block tree. Their approach is by obtaining domain level visual consistency by using visual features. Visual consistency can direct to high extraction accuracy even though template consistency may be missing. But this method is only compatible with the V-wrapper domain only.

On the other hand, the field study conduct by Yongquan Dong et al (2008) found a generic news extraction method that can easily identified news content based on a set of combined heuristics and to exact every part of news according to a predefined schema [16]. This approach does not depend on any template. It uses the common principles of news to identify the news articles and does not consider any concrete tags and structure. The drawback of this method is, still there are more modifications to be made on the algorithm so that it will be able to handle more types of error.

This is the sample of system layout from other application which I find have the similarities with my system.

Google News.

Google news is an automated news aggregator provided by Google Inc [17]. Google News provides searching, and the choice of sorting the results by date and time of publishing.

The screenshot shows the Google News interface for Malaysia. At the top, there is a search bar with 'Search News' and 'Search the Web' buttons. Below the search bar, the page is titled 'Google news Malaysia' and includes an 'Advanced news search' link. A navigation menu on the left lists categories like 'Malaysia', 'Southeast Asia', 'World', 'Business', 'Sci/Tech', 'Sports', 'Entertainment', 'Health', and 'Most Popular'. The main content area is divided into 'Top Stories' and 'In The News' sections. The 'Top Stories' section features several headlines with accompanying images and source information, such as 'PM: Inquest only if any suspicion over Ahmad Sarbani's death' from Malaysia Star and 'Opposition Is Cheap Glass Posing As Diamond, Says Taib' from Bernama. The 'In The News' section lists 'Sebastian Vettel' and 'Fernando Alonso' from the Malaysian Grand Prix.

Figure 2.3: Google news

CHAPTER 3

METHODOLOGY/PROJECT WORK

3.1 PROJECT METHODOLOGY

In many ways building this system required four fundamental phases which are planning, analysis, design and implementation. This chapter will focus more on the methodology used to build the system. This system in this case uses prototyping-based methodology which falls under Rapid Application Development category.

This project uses prototyping-based methodology which performs the analysis, design and implementation phase concurrently and all these three phases are performed repeatedly in a cycle until this system is completed [18]. With this methodology, the basics of analysis and design are performed, as for the system; I have done a lot of analysis based on other research related, and analysis on why the system is needed to help the user on getting update of the current issues. By using this methodology, the work immediately begins on a system prototype; a 'quick-and-dirty' program provides a minimal amount of features.

The first prototype is usually the first part of the system that the user will use. This is shown the users and the project sponsor who provide comments which are used to re-analyze, re-design, and re-implemented a second prototype that provides few more features. This process will continue until the prototype provides enough functionality and satisfied the user and it will be released and used. After the system is released, refinement occurs until it is accepted as the new system.

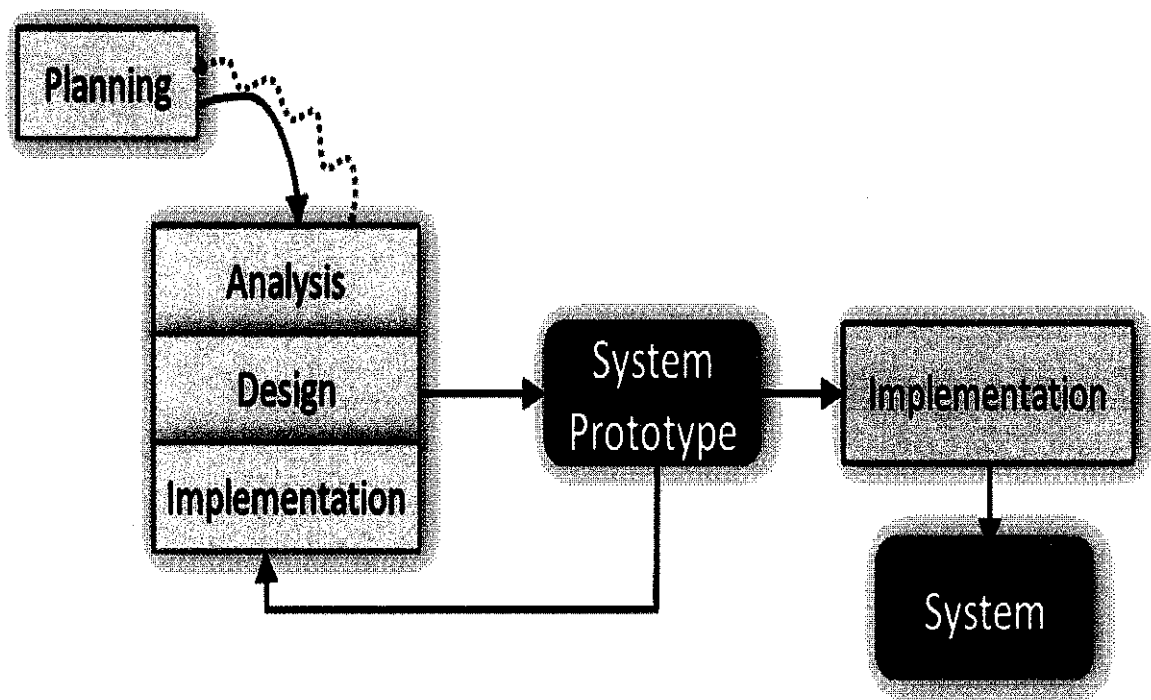


Figure 3.1: System methodology; prototyping-based

The reason of choosing this type of methodology is to let the system to be improved for a better system in a future with the comments from the users.

3.1.1 Planning

Planning phase is the fundamental process of understanding on why the project is should be built and determining on how the management process of the project would go [18]. In this phased, the key aspect of the project must be ask, “Can the system being build?”, “Will it give any benefit?” and lastly “If it was built, will it being used?”. Apart from that, in this stage the project title is determined and how the project will valued to the target user.

In this case, brainstorming and problem identification have been identified. From the problem that had been identified, the clear objective had been derived and the project title is proposed. Literature review had been conducted and the project activities and milestone (Gantt chart) had also been developed (See Appendices).

3.1.2 Analysis

Analysis phase is where the researches on current system, identify improvement from the current system and concepts on developing the system [18]. In the context of this project, research on the web data extraction system, techniques use, the type of method of extraction have been done. The understandings of the system have been congregated through the study of the journal and paper done on the web data extraction.

3.1.3 Design

The design phase decides how the system will operate, in terms of hardware, software and network infrastructure, the user interface, forms and the specific programs, databases and files that will be needed [18]. In other word, the steps in design phase determine exactly how the system will work. The framework and the architecture of the application have been done. The design phase will be the first phase of the iterative phase of the prototyping methodology.

3.1.4 Implementation

This is the final stage in the software life cycle before the first prototype is delivered. This is the most time-consuming phase, whereby all the programming, code-generating and the system improvement will be done to build the system as per user requirement. The prototype is build based on the deliverables in the design phase such as architecture design, interface design and database design. In the prototype based methodology, in this stage the ‘quick and dirty’ program that provides a minimal amount of features is delivered. The project will keep on growing or updating until there are enough features that have been implemented and after the review from the user.

After implementing the draft system to be use to the user in the real environment, users gave their feedback for improvement. In this project the system will extract the online news and display the headlines first.

3.1.5 System Testing

After the prototype is build, the system need to be tested with the user to make sure that the system is build as per user requirement. The user is the end-user/target audience who will be using the system. At this moment any faulty cause by the system will ne recorded for future use.

Basically, this phase is done iteratively with the implementation phase as the testing need to be conducted when the prototype is build or upgraded. The test will be conducted in the user environment to get the best result of the testing. Any errors occur will be recorded.

3.1.6 Final System

When the project reached this phase, there will be no more iterative process. In other word, this is the last phase where all the prototypes are built completely and system is upgraded based on user feedbacks. The real product/ the final system will be delivered and launched at the end of the project milestone. And there will still be maintenance to keep the system reliable with the current situation.

As for final year project, this project is developed up until first prototype of the system. For this system, the first prototype is aims to deliver the function of extracting the data/ information as determined.

3.2 Proposed Framework

Project framework is the fundamentals of project management which composed of nine knowledge areas: project integration management, project scope management, project time management, project cost management, project quality management, project human management, project communication management, project risk management, project procurement management and five processes: initiation, planning, executing, controlling, and closing [19]. Below is the project framework

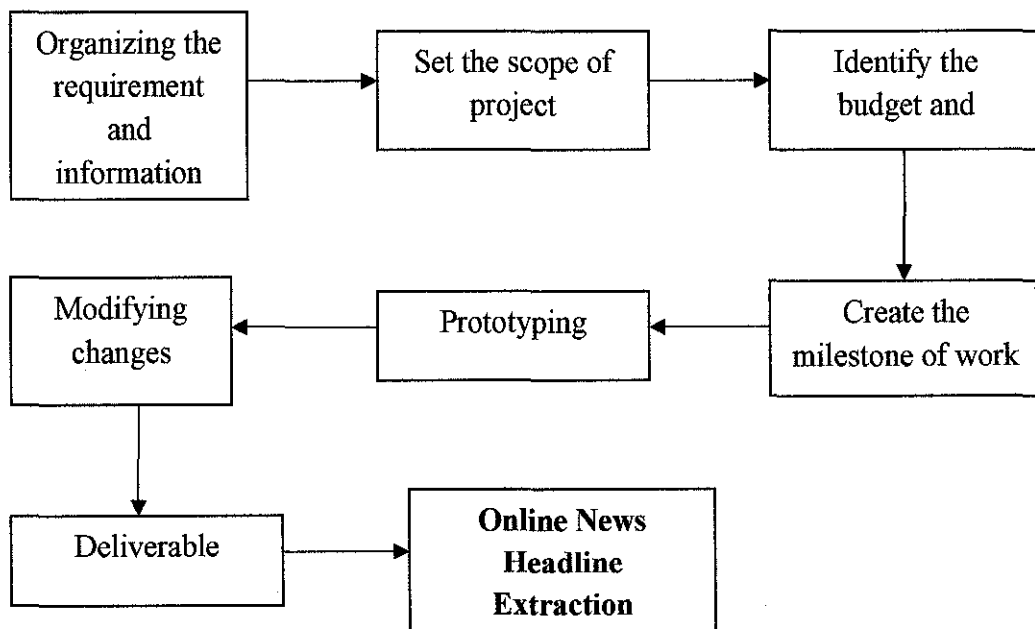


Figure 3.2 : Proposed framework

Based on the figure 3.2 above, first thing have to be done is organizing the requirement and the information on the system. The requirement needed for this system is to make sure that the system is able to extract the information on the online news and to make sure that the news is auto updated.

Second is, to set the scope of project. To cope with the time this project will extract the news headlines from the local online newspaper. The news extracted will be from the top news from the New Straits Times, The Star and Malay Mail. Next is, to identify the budget and resources, this is to make sure that the system can be done within the budget given and the resource that available.

After that, create the milestone or the Gantt chart of the project work. This is to make sure that the project is well planned and can be finished in the time given. And then, the first prototype will be delivered. Based on the review done by the user, some changes will be done and the system will be modified before the real system is being delivered.

CHAPTER 4

RESULT AND DISCUSSION

4.1 System Architecture

This is the proposed E-Headlines news system structure for the Online News Headline Extraction project.

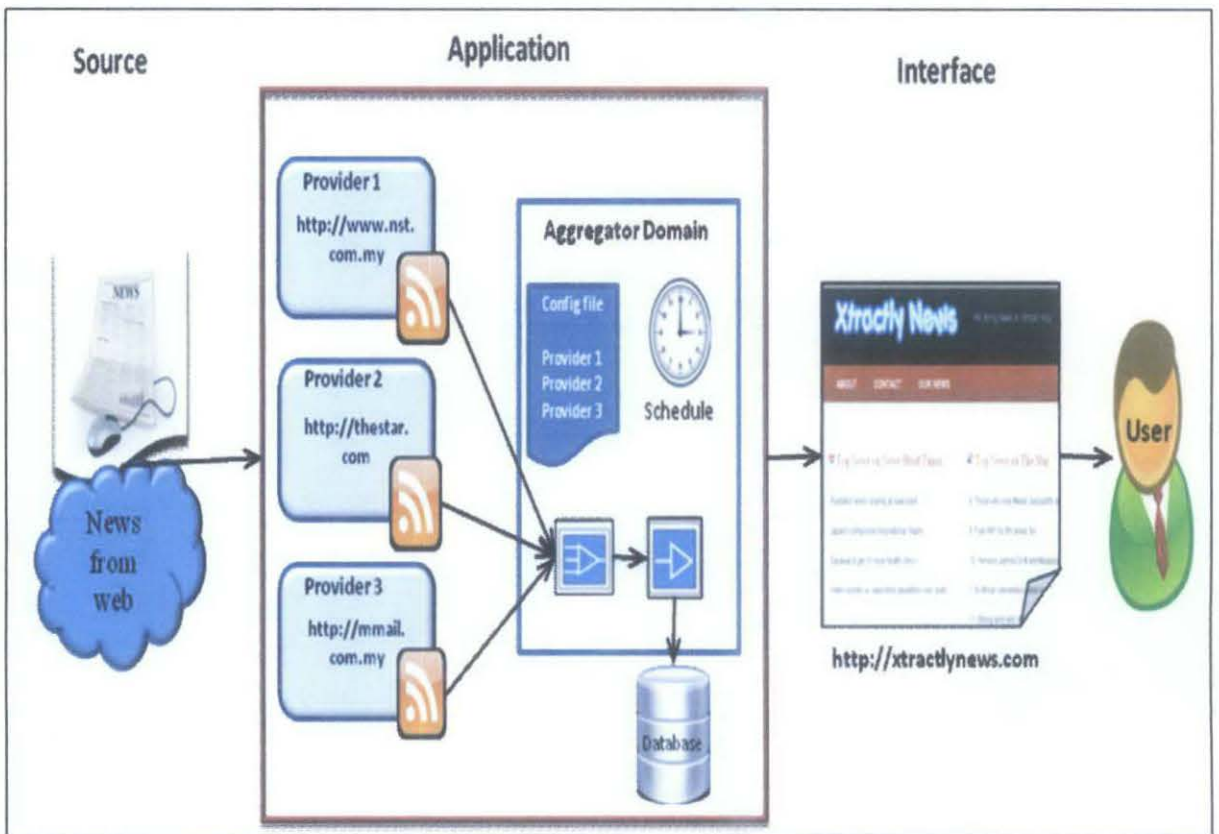


Figure 4.1: System Architecture

The figure 4.1 above shows the system architecture of the system. The sources of the system are from the news on the web. In the application of the system, there will be three main providers that are The Star, New Straits Times and Malay Mail. The aggregator or also own as feeds reader will subscribe the RSS feed of the top news from this three website. The aggregator will combine all the feeds to be put on a websites and augment it into the database. For a conclusion, this application will extract the headlines from the news, put it on the website and display it in the manner that satisfied the user need.

4.2 System Process

Below are all the steps and process that involved in building this system

1. Buy the domain and the hosting



Figure 4.2.1: cPanel

2. Login cPanel

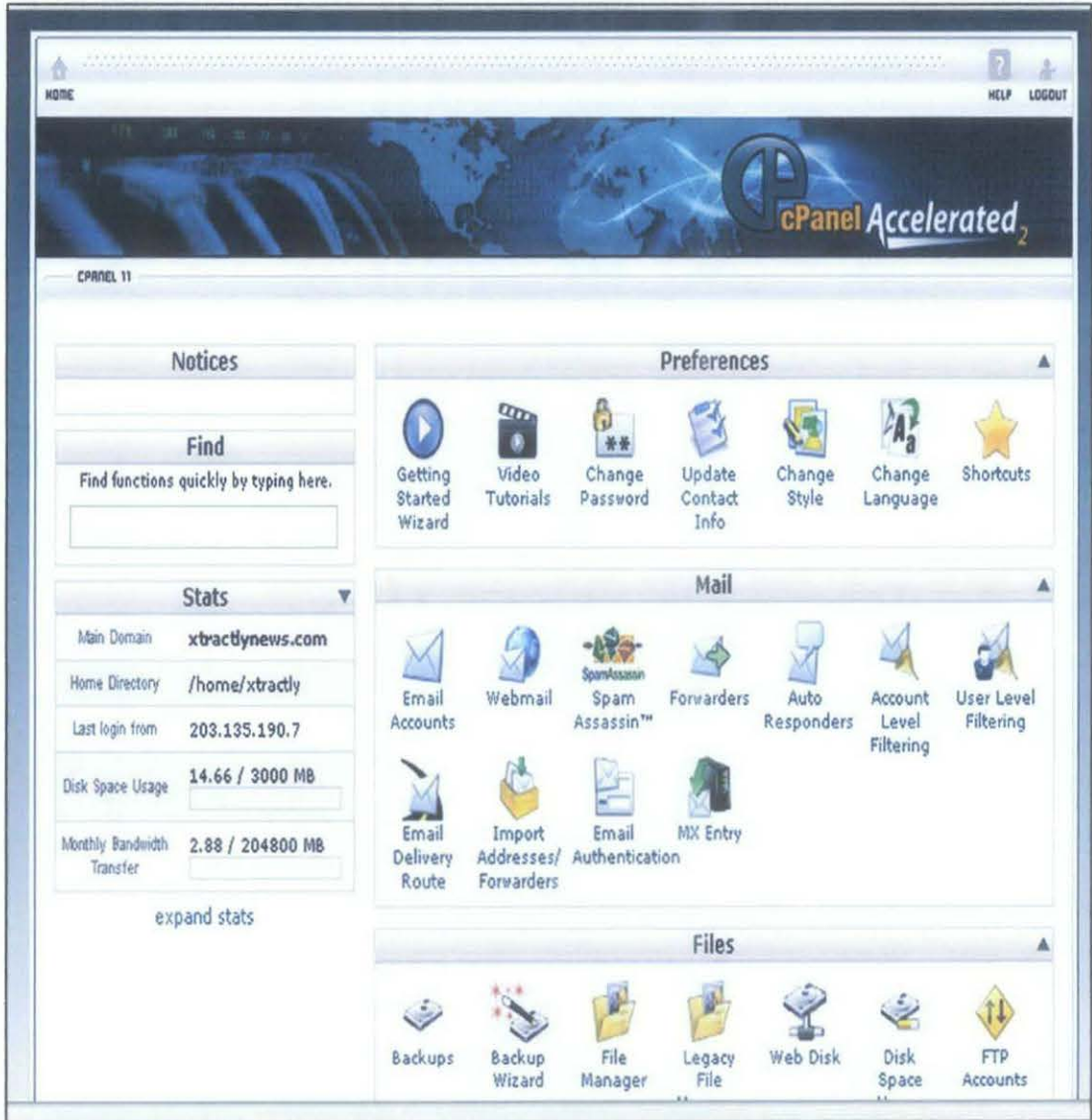


Figure 4.2.2: cPanel Homepage

3. Create database in the cPanel. The database is for the system's website.

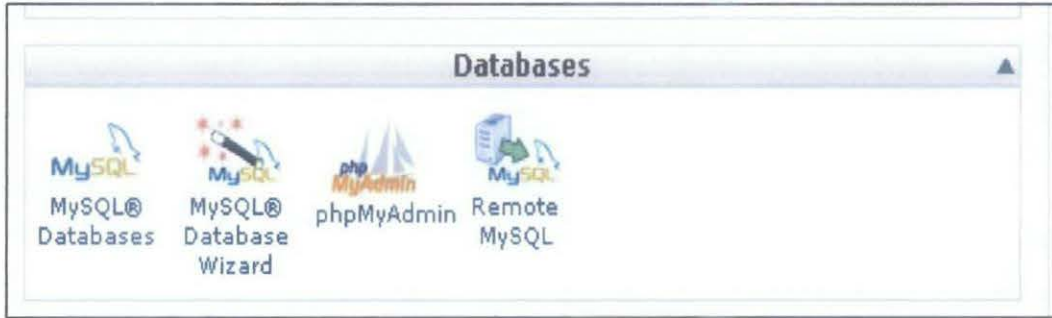


Figure 4.2.3.1 : MySql Database

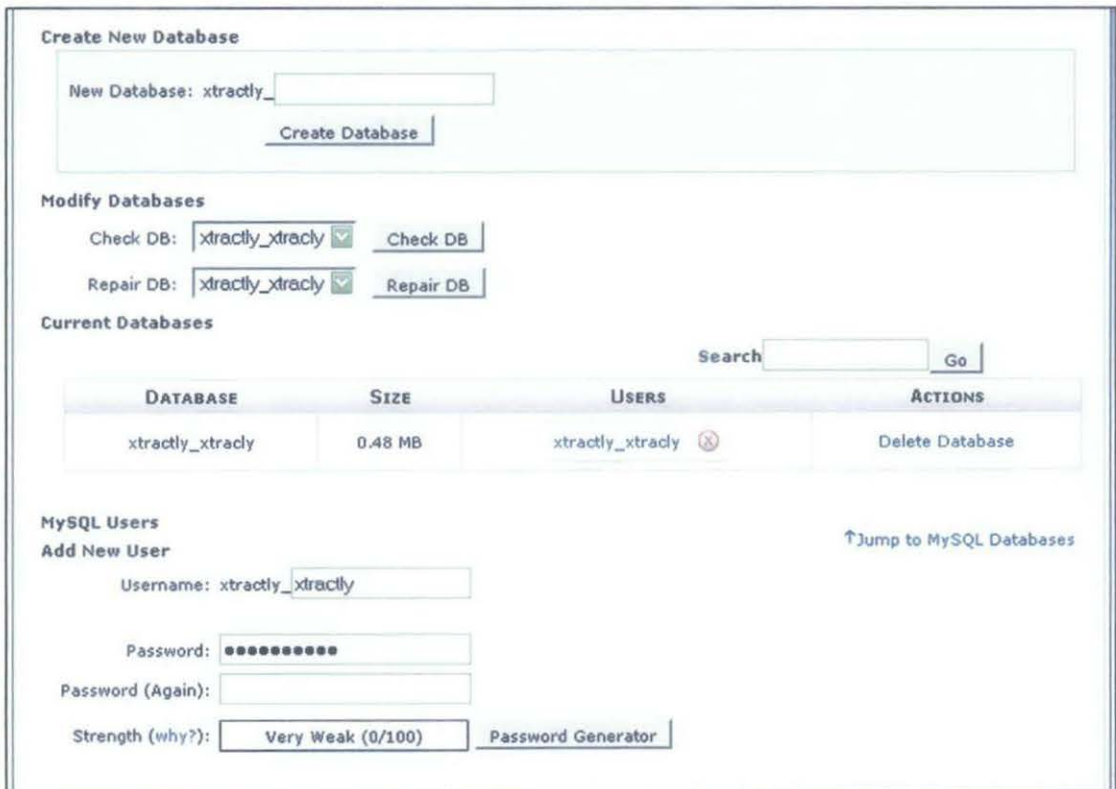


Figure 4.2.3.2 : Current Database

In phpMyAdmin, the database create are for the table for receiving and publishing the news on the website.

The screenshot shows the phpMyAdmin interface for a database named 'xtractly_xtracly' on localhost. The 'Structure' tab is active, displaying a table of database tables. The table has columns for 'Table', 'Action', 'Records', 'Type', 'Collation', 'Size', and 'Overhead'. The tables listed include wp_commentmeta, wp_comments, wp_contact_form_7, wp_links, wp_options, wp_postmeta, wp_posts, wp_terms, wp_term_relationships, wp_term_taxonomy, wp_usermeta, and wp_users. A summary row at the bottom indicates 12 tables with a total of 283 records and a size of 492.2 KiB.

Table	Action	Records ¹	Type	Collation	Size	Overhead
<input type="checkbox"/> wp_commentmeta		4	MyISAM	utf8_general_ci	7.2 KiB	-
<input type="checkbox"/> wp_comments		31	MyISAM	utf8_general_ci	21.1 KiB	-
<input type="checkbox"/> wp_contact_form_7		1	MyISAM	utf8_general_ci	3.6 KiB	-
<input type="checkbox"/> wp_links		7	MyISAM	utf8_general_ci	3.5 KiB	-
<input type="checkbox"/> wp_options		155	MyISAM	utf8_general_ci	366.8 KiB	5.5 KiB
<input type="checkbox"/> wp_postmeta		25	MyISAM	utf8_general_ci	10.1 KiB	-
<input type="checkbox"/> wp_posts		24	MyISAM	utf8_general_ci	52.2 KiB	36 B
<input type="checkbox"/> wp_terms		2	MyISAM	utf8_general_ci	8.1 KiB	-
<input type="checkbox"/> wp_term_relationships		11	MyISAM	utf8_general_ci	3.2 KiB	-
<input type="checkbox"/> wp_term_taxonomy		2	MyISAM	utf8_general_ci	4.1 KiB	-
<input type="checkbox"/> wp_usermeta		20	MyISAM	utf8_general_ci	8.1 KiB	-
<input type="checkbox"/> wp_users		1	MyISAM	utf8_general_ci	4.1 KiB	-
12 table(s)	Sum	283	MyISAM	latin1_swedish_ci	492.2 KiB	5.5 KiB

Check All / Uncheck All / Check tables having overhead With selected:

Print view Data Dictionary

Create new table on database xtractly_xtracly

Name: Number of fields:

Figure 4.2.3.3 : phpMyAdmin Database

4. Get the RSS from the online news website.

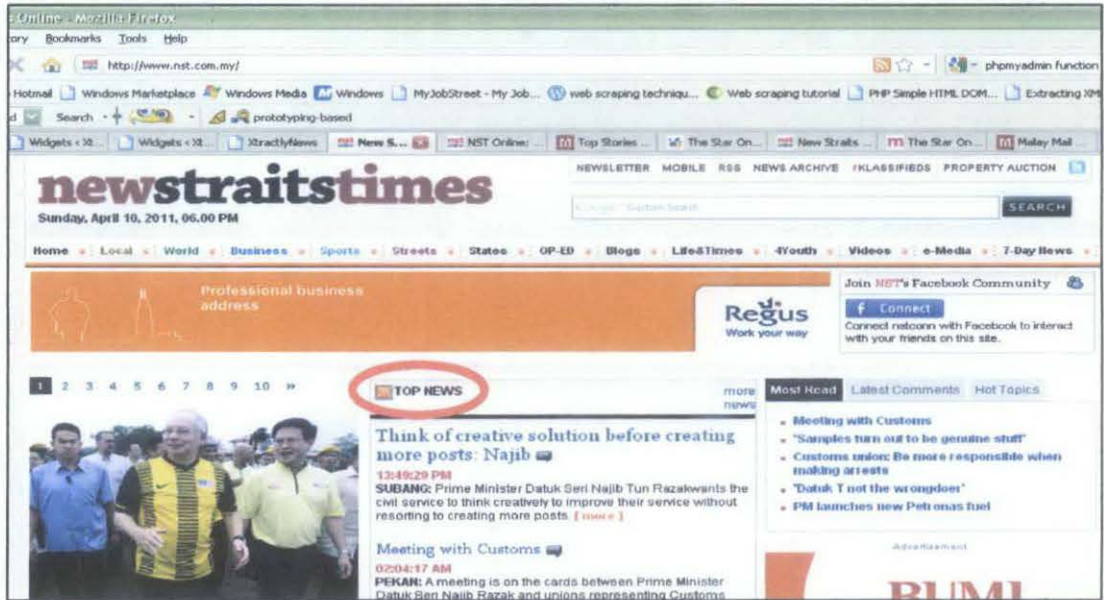


Figure 4.2.4.1: New Straits Times online [20]

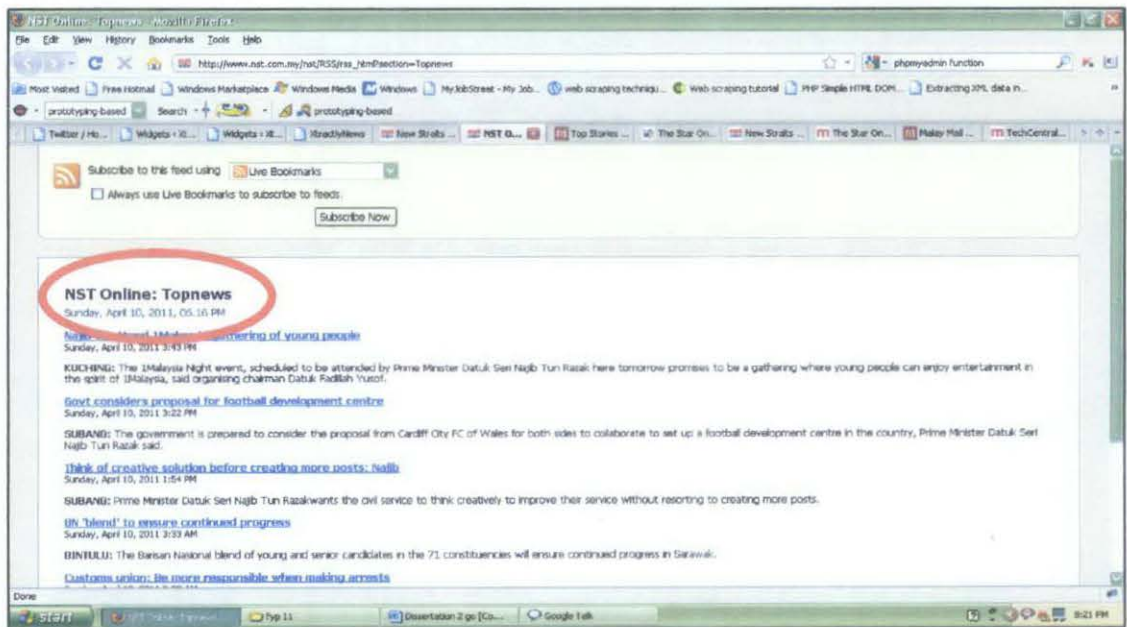


Figure 4.2.4.2: Top news on New Strait Times

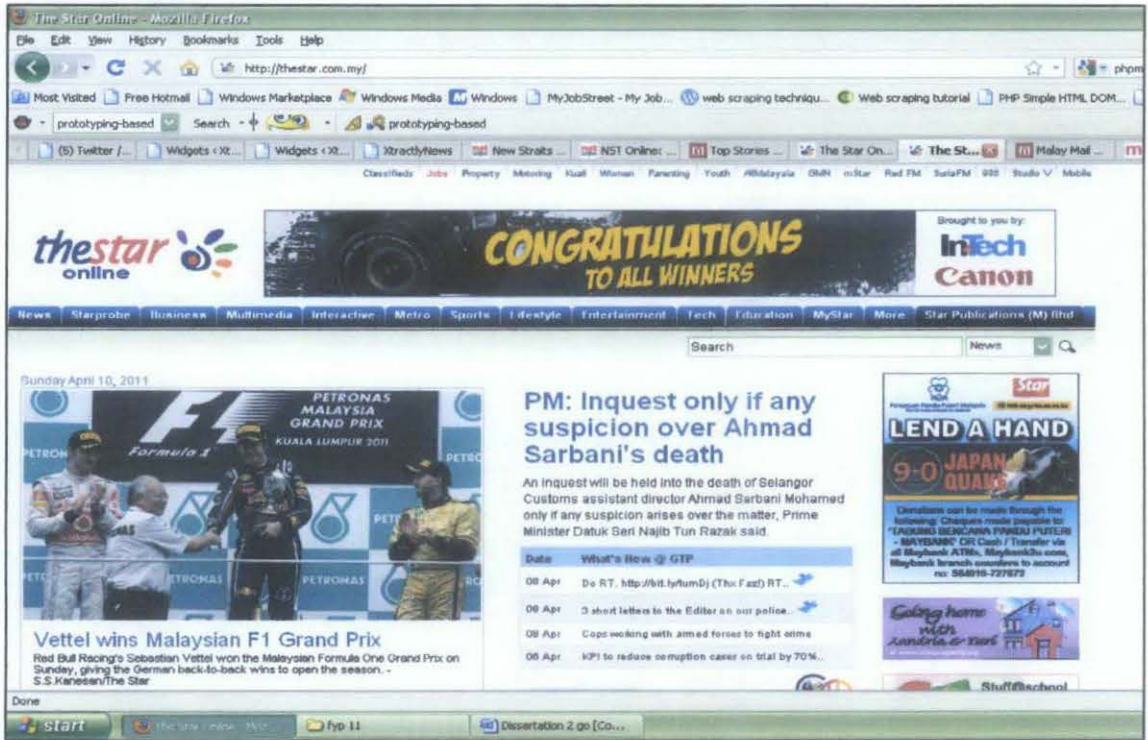


Figure 4.2.4.3: The Star online [21]

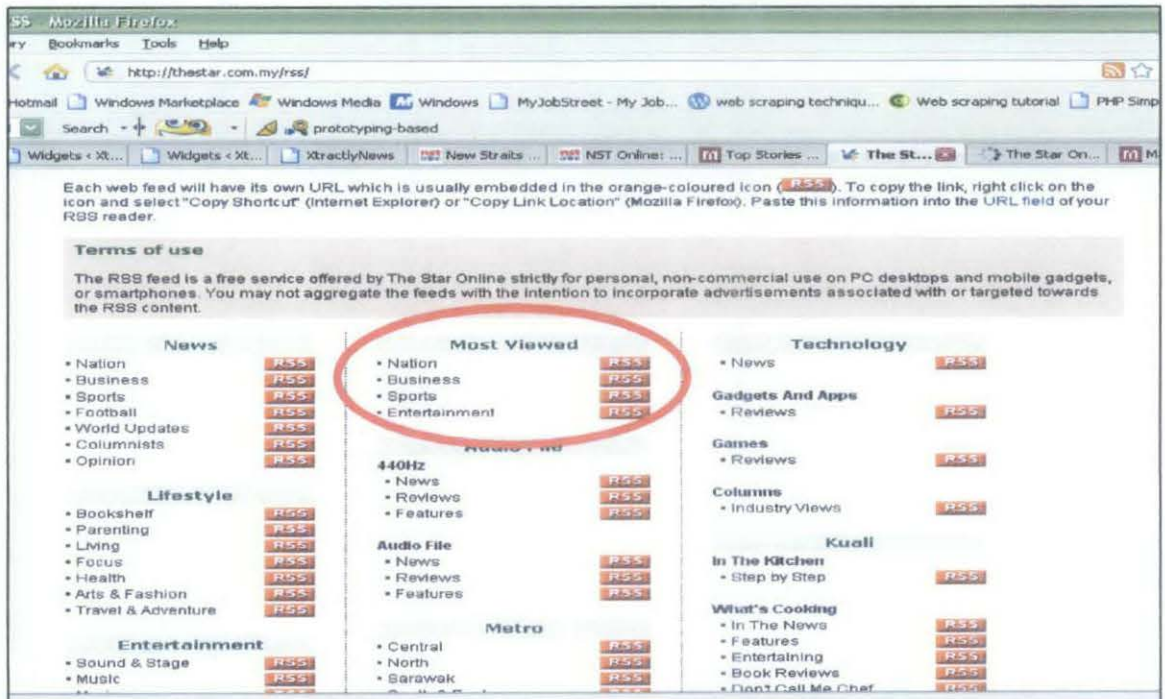


Figure 4.2.4.4: Top news on The Star



Figure 4.2.4.5: Malay Mail online [22]

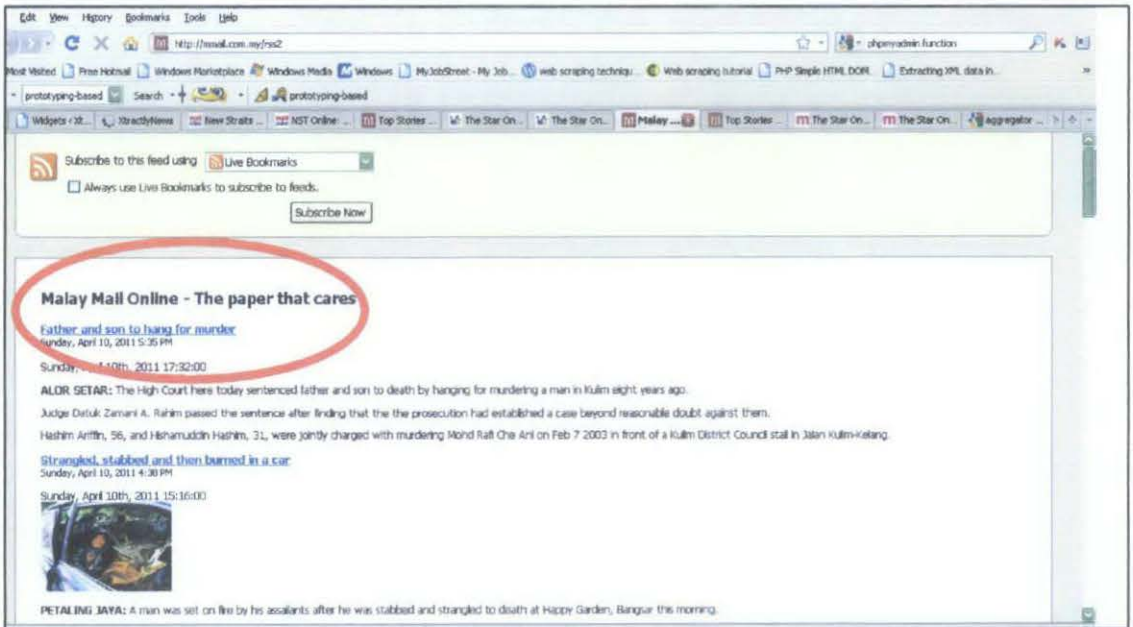
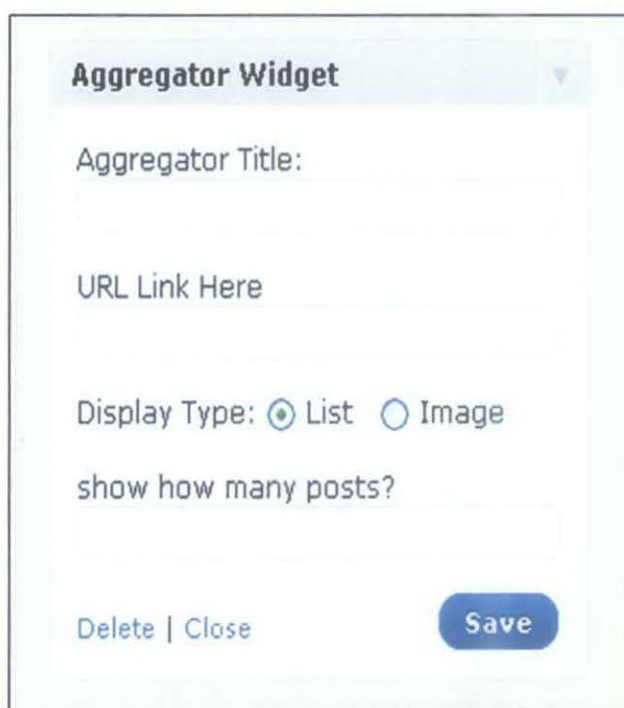


Figure 4.2.4.6: Top news on Malay Mail

5. Paste the feed URL (RSS) into the aggregator.



The image shows a configuration window for an "Aggregator Widget". The window has a title bar with the text "Aggregator Widget" and a small downward arrow on the right. Below the title bar, there are several input fields and controls:

- A text input field labeled "Aggregator Title:".
- A text input field labeled "URL Link Here".
- A "Display Type:" section with two radio buttons: "List" (which is selected) and "Image".
- A text input field labeled "show how many posts?".
- At the bottom left, there is a link that says "Delete | Close".
- At the bottom right, there is a blue button labeled "Save".

Figure 4.2.5.1: Aggregator widget

Write the title of the aggregator into aggregator title and paste the URL of the RSS feed at the home page feeds. This is the news that will be displayed on the homepage of the website.

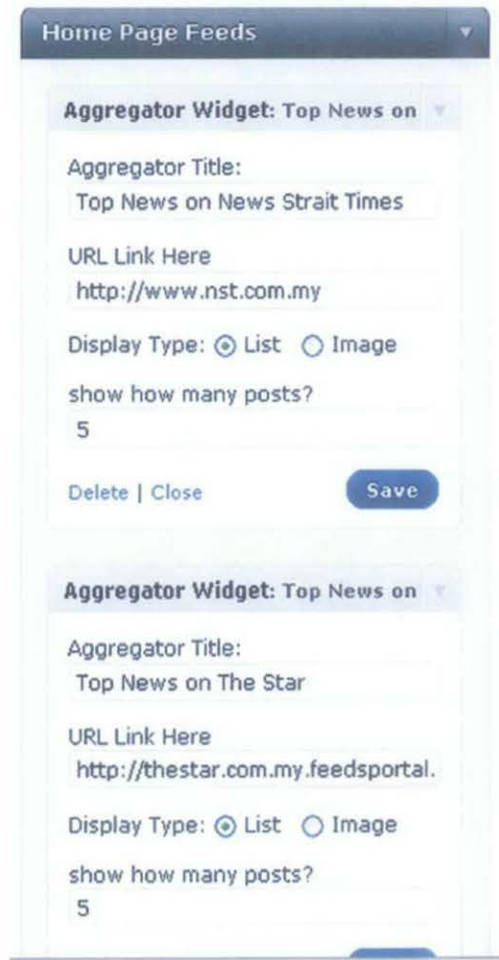


Figure 4.2.5.2: Home page feeds

6. Interface of the system.

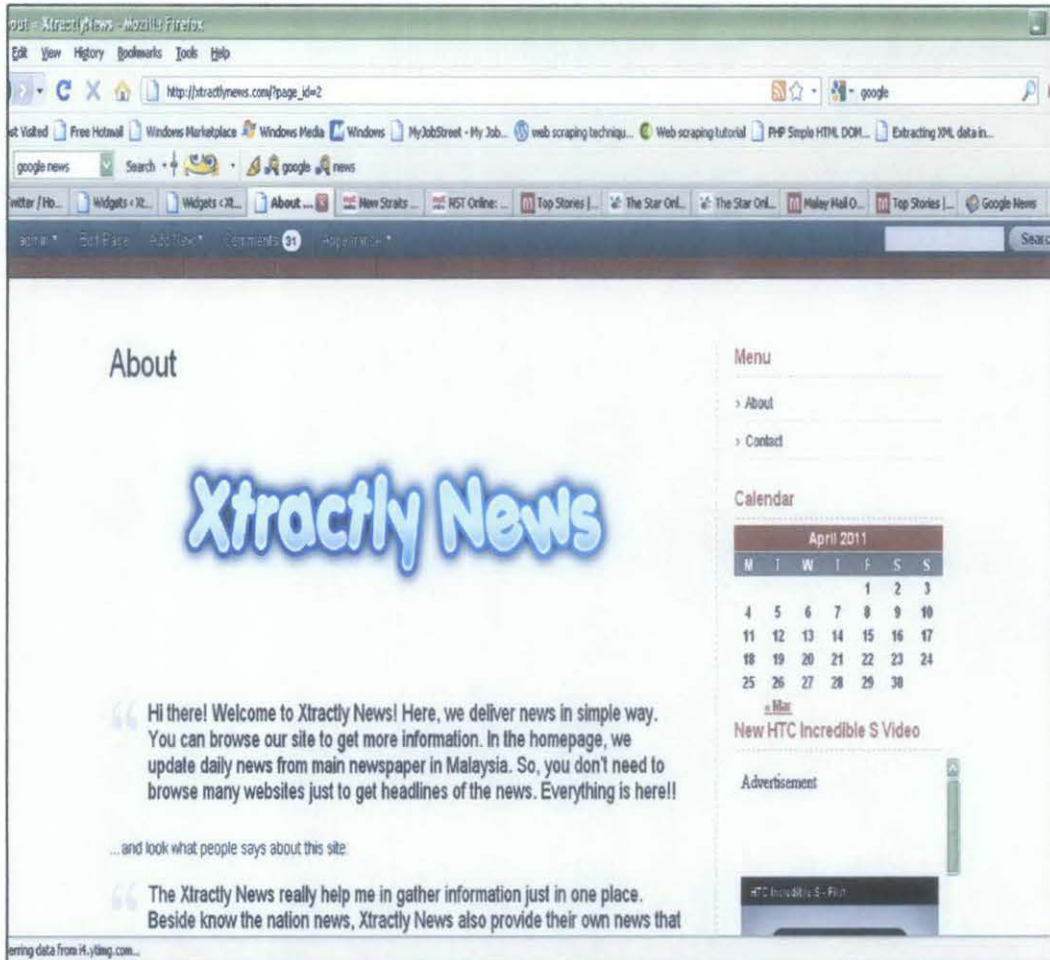


Figure 4.2.6.1: About Xtractly News

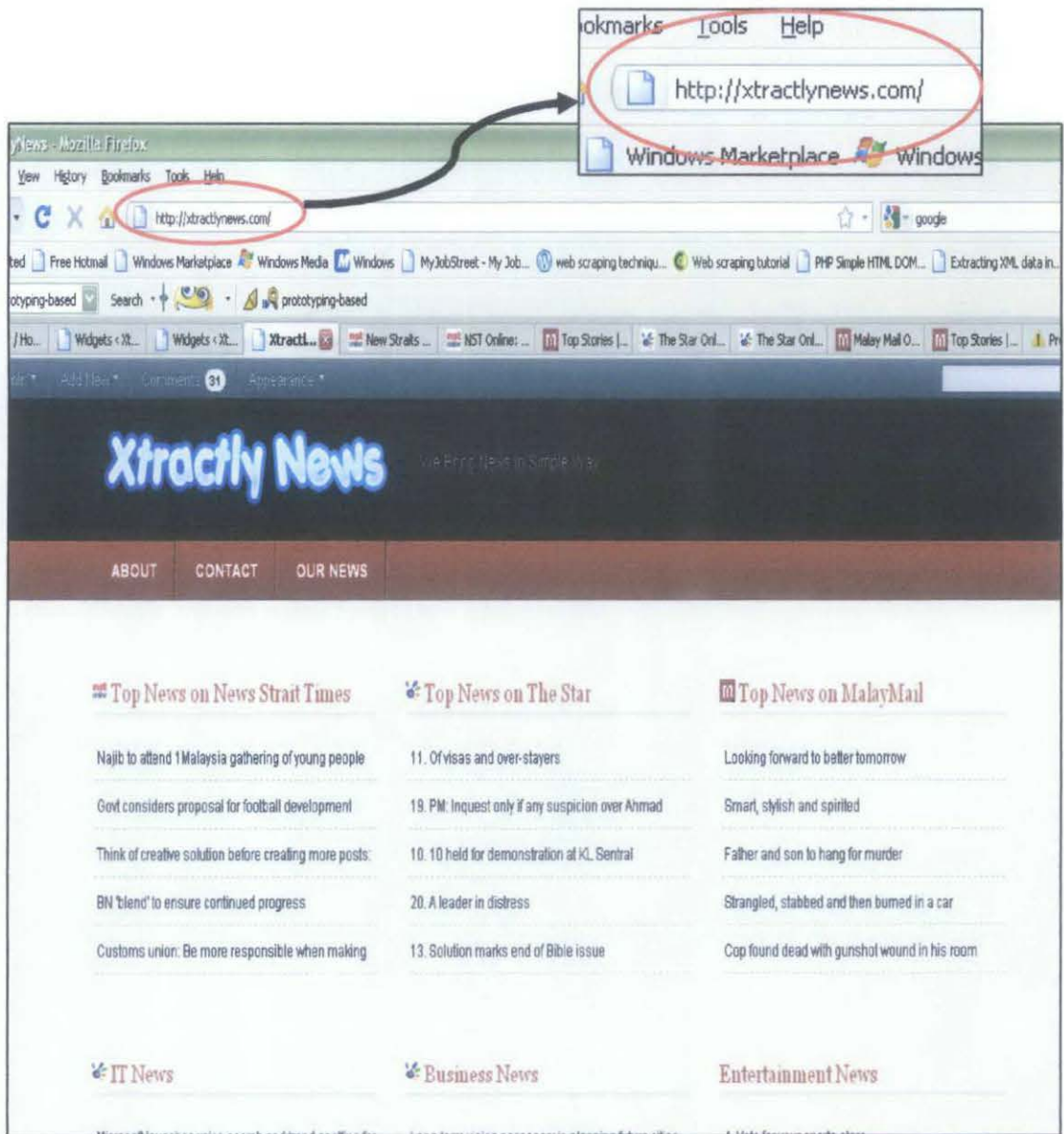


Figure 4.2.6.2: Snapshot of the interface

CHAPTER 5

CONCLUSION AND RECOMMENDATION

This project highlights on the online news headline extraction that can be useful to the user. The main contribution of this project is the proposal of an E-Headlines News Extraction Framework that illustrated the extracted information on the news. This system will help user overcome the information overloading that encounter the net nowadays. By having this system, the user can retrieve the news that they are interested in faster without wasting so much time on it. This system also supports the green-technology by reducing the browsing time and eventually save the energy.

In future, this system can be enhanced by adding more extracted information on the news such as the summarization of the news. This system also can be upgrade by having a better interface.

REFERENCES

- [1] Miniwatts Marking Group, <http://www.internetworldstats.com/>

- [2] Cunningham, H., November 2004. Information Extraction, Automatic. Department of Computer Science, University of Sheffield.

- [3] V. Pinheiro, T. Pequeno, V. Furtado and D. Nogueira. Information Extraction from Text Based on Semantic Inferentialism. Departamento de Ciencias da Computacao Federal University of Ceara (UFC), Fortaleza, Ceara, Brazil, Mestrado em Informatica Aplicada University of Fortaleza (UNIFOR) and ETICE Fortaleza, Ceara, Brazil.

- [4] Man I.L., Zhiguo G., and Maybin M., 2008. A Method for Web Information Extraction, Faculty of Science and Technology, University of Macau, Macao, PRC

- [5] Tracy Y. and Suzy B., 2005. Online Newspapers Enjoy Double Digit Year-Over-Year Growth, Reaching One Out Of Four Internet Users, According To Nielsen//NetRatings.

- [6] D Applet, J Hobbs, J Bear, D Israel, M Kameyana, and M Tyson, 1993. Fastus: a finite- state processor for information extraction from real-world text.

- [7] J. Cowie and W. Lenhart, 1996. Information extraction. Communications of the ACM, 39(1): 80-91.

- [8] P Jackson and I Moulinier, 2002. Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization. John Benjamins, Amsterdam.
- [9] Norshuhanani Zamin, 2009. Information Extraction for Counter-Terrorism: A Survey. Department of Computer and Information Sciences, Universiti Teknologi PETRONAS.
- [10] D Robert, P Cecile and T Marc, 2003. Information Extraction Via Path Merging. Centre for Language Technology, Macquarie University, Sydney and Intelligent Interactive Technology Group, CSIRO, Sydney.
- [11] B. Fazinnga, S. Flesca, and A. Tagarelli, 2009. Schema-based Web wrapping, Department of Electronics, Computer and System Sciences, University of Calabria, CS, Italy.
- [12] Hua W. and Yang Z., 2010. Web Data Extraction Based on Simple Tree Matching, College of Information Engineering, Northwest A&F University, Yangling, China.
- [13] Hristo T., Jakub P., and Martin A., 2008. Real Time News Event Extraction for Global Crisis Monitoring, Web and Language Technology Group of IPSC, Ispra, Italy.
- [14] Jakub P., Hristo T., Martin A. and Erik V., 2008. Cluster-Centric Approach to News Event Extraction, Institute for the Protection and Security of the Citizen, Ispra, Italy.
- [15] Shuyi Z., Ruihua S., and Ji-Rong W., 2007. Template-Independent News Extraction Based on Visual Consistency, Pennsylvania State University, University Park, and Microsoft Research Asia, Beijing, China.

- [16] Yongquan D., Qingzhong L., Zhongmin Y., and Yanhui D., 2008. A Generic Web News Extraction Approach, School of Computer Science and Technology, Shandong University, Jinan, P.R. China and School of Computer Science and Technology, Xuzhou Normal University, Xuzhou, P.R. China.
- [17] Google News, <http://news.google.com/>
- [18] Alan D., Barbara H.W. and David T., 2005. System Analysis and Design with UML Version 2.0, An Object-Oriented Approach, Second Edition, (p4-12).
- [19] Definition of Project Framework. <http://sites.google.com/site/pmpbank/glossary>
- [20] New Straits Times, <http://www.nst.com.my/>
- [21] The Star, <http://thestar.com.my/>
- [22] Malay Mail, <http://mmail.com.my/>

Appendices

Appendix 1-1: Gant Chart of the project

(First Phase)

Task/ Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Identification of problem														
System proposal														
Research and literature review														
Preliminary report submission				x										
Research and literature review														
Analysis on the requirement														
Submission of progress report								x						
Developing the architecture of the system														
Interim report submission													x	
Developing the Graphical User Interface (GUI)														
Project presentation														x

	Process
x	Milestone

(Second Phase)

Task/ Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Study on system architecture															
Developing the interface for the system															
Building the prototype															
System Improvement															
Poster exhibition										X					
Finalize on the system															
Final presentation														X	
Final report submission															X

	Process
x	Milestone

Appendix 1-2 : Project Activities

