IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 4, Ver. I (Jul.-Aug. 2016), PP 59-65
www.iosrjournals.org

Design and Development of an Automatic Online Newspaper Archiving System

Md. Ashik Saeed[1], Tanzina Rahman[1]
[1](*Department of Applied Physics and Electronic Engineering, University of Rajshahi, Bangladesh)*

Abstract: *News archive has always been a great source of information. Till date, several printed and manual information retrieval systems served the purposes. But these systems are gradually becoming out of date, since they offer limited scope and facility for searching and retrieval of information and require a large amount of storage capacity. This study aims to illuminate a prototype of an automatic online newspaper archiving and information retrieval system based on online newspaper websites, which will offer a large amount of information storage capacity and easy navigation system for users. In this research work we conducted an inductive study on the structure of some mostly visited online newspaper sites and at the same time an effort was made to develop an algorithm for a news crawling program that will make it possible to store news from online news sites automatically with minimum bandwidth wastage and provide an easy User Interface for users to browse and search for news. The findings from the study suggest that, in perspective of Bangladesh no systematic structure is used for the development of online news websites. These websites can be brought under a standard structural format which will make archiving easier and error free. Finally we had developed an automatic news archive system based on a web crawler program and propose some structural considerations for online news sites that will make the site friendly for archiving.*

Keywords: *Archiving system, database design, User Interface design, Java programming language, Web crawler*

I. Introduction

One of the most precious gift human acquired through the advancement in technology is newspaper. In this age of information newspaper is a powerful tool to express ideas, information and to create awareness among the people. Ever since the formation of society, newspaper has always helped people. It forms an essential part of modern civilization. Newspaper always has been a source of great amount of information. The increasing need and demand for use of newspaper information, from among the people involved both with the research and development activities, and daily user can be significantly made easy with a news archive. An archive is an accumulation of historical records. In general, an archive consists of records that have been selected for permanent or long-term preservation. Archives contain primary source documents that have accumulated over the course of an individual or organization's lifetime, and are kept to show the function of that person or organization [1].

A news archive keeps a collection of newspaper for people to access for a long period. A news archive thus unlock doors to the past for historians, genealogists, lawyers, demographers, filmmakers, student and others to conduct research, by serving them access to previous day's news. As the internet access is spreading rapidly day by day, online newspaper website has become more popular for information seeker. People from different aspect of life feel free to use online newspaper more than a printed newspaper for its easy accessibility and mobility. From a study over the most visited sites from our country we can see that about 30% – 35% are online news sites [1]. Today whenever anyone needs to know or find something, they feel it easier to search through the internet, since it gives them a vast amount of information and recommendation. It will be therefore more reasonable to develop a news archive based on online newspaper. Although these websites have their own archive, but almost all of their archives have limited capacity, do not have issues more than five or six years old. Proper archiving of this online newspaper can be of great benefit for our upcoming generations.

The primary goal of this research work is to develop a news archive based on the online news site which will automatically store news from a predefined set of online news site and introduce an easy navigation system for the user to find or search his desired news. As the online news website does not follow or have any standard structure or design rules to make the site archiving friendly, it is difficult to archive these news sites without any information loss.

The second goal of this research work is the archiving system to point the problems and feature of online newspaper sites that will help to develop and propose an idea about the structure of these news sites that will make it easier for organizations to archive these news sites with minimal information loss.

This research work made it possible to develop an automatic news archiving system that can navigate through an online newspaper site looking for published news and save pages containing news. This is made

possible by using a web crawling program. The archiving system also enables users to browse for any previous day's version of desired newspaper site easily and search for news from different news site that helps the user to find the news from different source in a same place.

This study enhances our understanding about the structure and different feature of online newspaper website. This work explores for any consideration that will help make an online newspaper website archiving friendly and enable organizations to archive these news site with more ease and error free.

II. web crawler and system architecture

A Web Crawler also known as "Web Spider", "Web Robot" or simply "bot" is software for saving pages from the web automatically. Unlike what the name may suggest, a Web Crawler does not actually move. around computers but only sends requests for document on the web servers from a set of pre-defined locations.

The input to this software is a set of pre-defined or seed URLs. Once the pages from that URLs or links are downloaded the pages are parsed and scanned for more links. The latest links pointing to pages that have not yet been downloaded, are added to a queue which are downloaded later and scanned for new links [2]. The process is repeated until a stop criterion is met.

In the simplest form a crawler starts from a seed page and then uses the external links within it to attend to other pages. The process repeats with the new pages offering more external links to follow, until a sufficient number of pages are identified or some higher level objective is achieved. Behind this simple description lies a host of issues related to network connections and parsing of fetched HTML pages to find new URL links.

A Web crawler is a tool for exploring a subset of the Web. This exploration may serve diff erent goals. The key application of a Web crawler is to create an index covering broad topics (general web search) or specific topics (vertical web search). Web crawlers may also be used to analyze Web sites automatically for extracting aggregate statistics (web characterization), for improving the sites (web site analysis), or for keeping a copy of a set of pages (web mirroring).

A Web crawler will perform the navigation through the website finding unsaved pages. The crawler will perform necessary actions to save any unsaved pages and its content to our local server, and process data while saving the page. A Database will act as the backbone of our archive's indexing system. It will contain the necessary information to keep track of the saved page such as the website, date of saving the page and other valuable data [1].Finally a User Interface will make it possible for users to browse and search for desired news.

1.1. Algorithm for Web Crawling Program

There is an intention to keep the algorithm for the web crawling program as simple as possible. The program starts with a given seed URL which should be a link to the homepage of our desired news website. The program saves the page and analyzes it for all the links contained by the page, extracts them and saves them in a queue. While saving all the links in a queue the program looks for any duplicate link and if found any, it discards them so that the queue contains only unique URLs [2, 3]. This first queue of unique URLs is called "level-1 URLs".

Now the program takes the first URL from the queue as input. Then the program tests if the page represented by the URL is saved before. To do so it takes help of the database containing the list of all URLs which have been saved before. If the page is not saved in previously it saves the page, adds the URL in database containing the list of all URLs, analyze it for all the links contained, extracts them and saves only the unique URLs in the queue. These new URLs obtained from any level-1 URL are called level-2 URL. In this process the program tests all level-1 URLs and creates a queue of level-2 URLs. Then it tests all level-2 URLs and creates a queue of level-3 URLs and so on. The program keeps saving any unsaved pages while navigating through higher level URLs. The program can be terminated by setting maximum URL level. The flow chart for Web Crawling Program is sown in the Fig. 1.

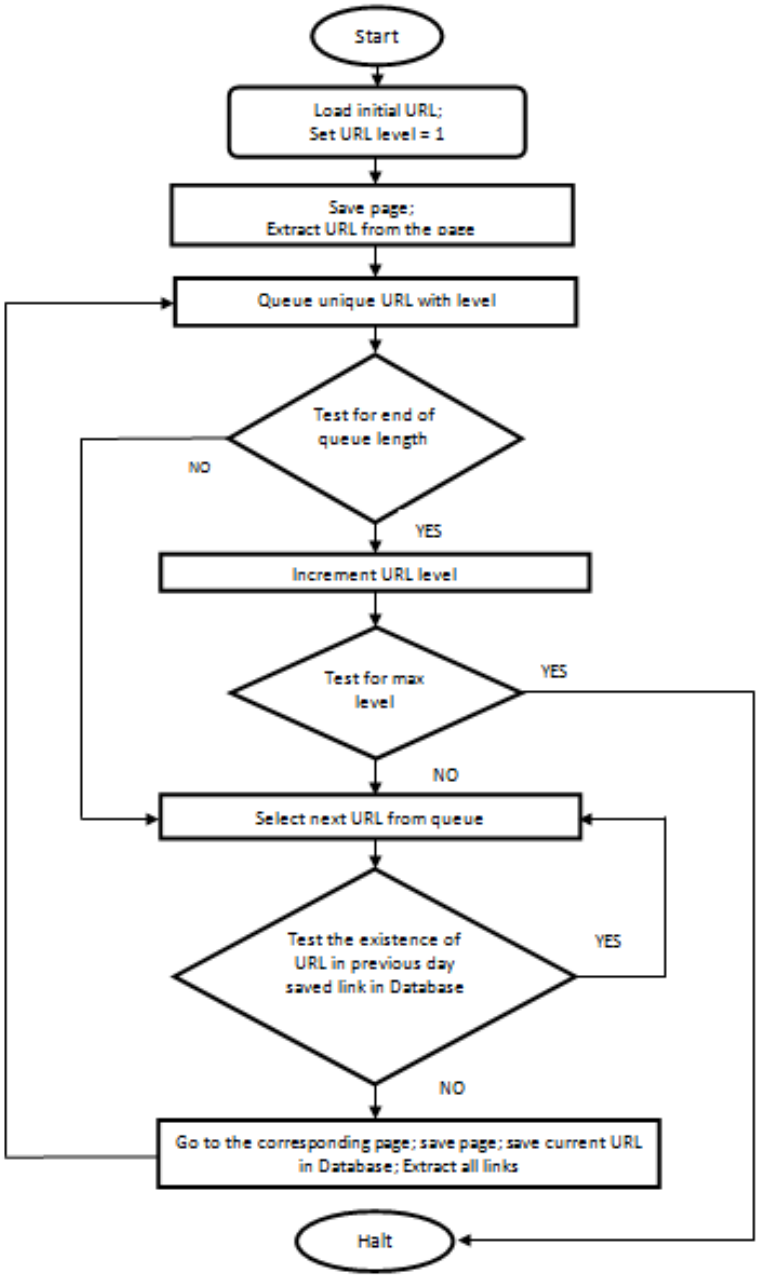


Fig. 1: The flow chart for Web Crawling Program.

1.2. Java as the Programming Language

In this research work the web crawling program is developed using java programming language since Java is easy to use, write, compile, debug, and learn than other programming languages. Because of Java's robustness, ease of use, cross-platform capabilities and security features, it has become the choice to develop the crawling program in java language [4].

Besides javas core library we have used a third party java library named jsoup which provides a very convenient API for working with real-world HTML, extracting and manipulating data, using the best of DOM, CSS, and jquery-like methods [18]. It also supports http compression during document parsing from the web which minimizes the bandwidth required for our crawling program.

Another third party java library used here is Apache Commons IO. It provides a multitude of classes that enables us to do common tasks such as saving and copying files easily and with much less boiler-plate code, that needs to be written over and over again for every single project [4].

1.3. Database Design

The database is certainly a very important section for our archiving system. Indexing system for saved page is fully dependent on the database in the archiving system. The database will also contain the links for the

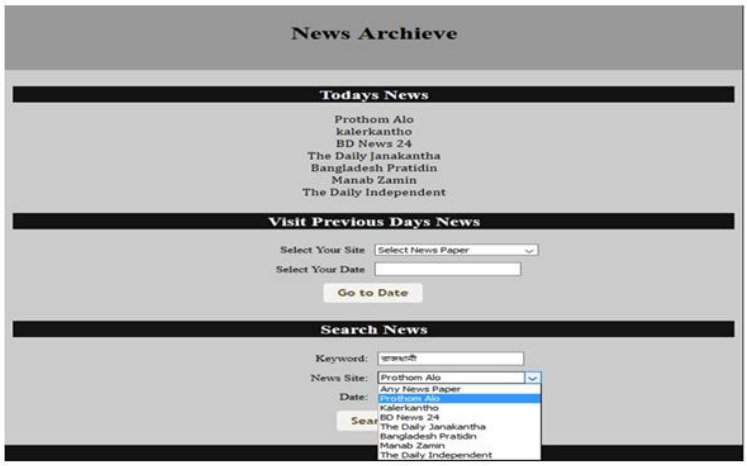
page which is saved in any previous day. The crawling program repeatedly communicates with the database during execution. A much simpler database is designed for the indexing of the archiving system. It contains several tables, each for a certain website containing several information about the saved page regarding to that website. Table 1 shows the fields of each table with their data type.

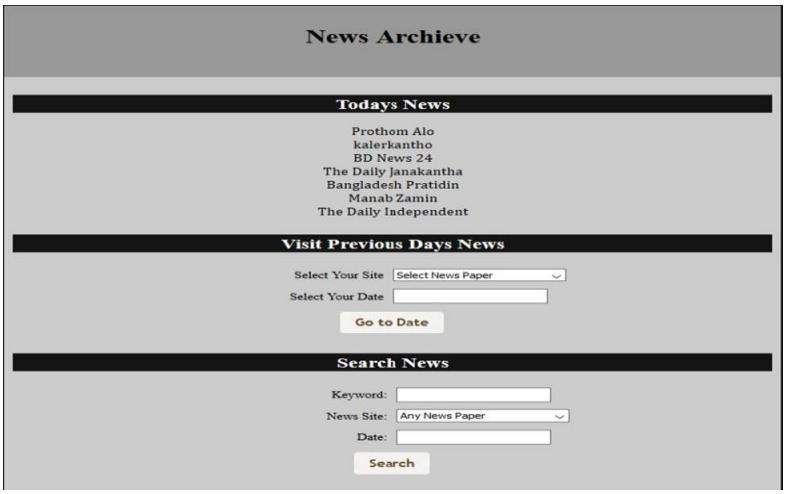
Field name	**Field Type**	**Maximum. Field Size**
ID	Integer	10
WebLink	VarChar	500
LocalLink	VarChar	500

The information of each link saved in the database table has a unique id number which is contained in the ID field of the table. Its type is integer, since it contains number of each link and its maximum size is 11 digits.The Web Link field contains the absolute URLs which are extracted from each page and its type is VarChar since it contains different type of character, number and symbol. The maximum size this field can contain is 500 digits.The last field is Local Link which contains the modified link of each URL or the address of that webpage in the local server of the archiving system. Its type is also VarChar and maximum size is 500 digits.

1.4. User-Interface Design

An easy and simple user-interface is designed for users to navigate through the archive and search for desired news without any complexity. The UI would provide users access to the current day's news as well as news from any other day and choose his desired news website. User can also search news based on his desired keywords, news publishing day or specific news website. The UI is designed with the help of HTML, CSS, JavaScript and PHP scripting language since they are open source and can be used to design interface easily [1,2]. Some images showing the design of the user interface is given in the Fig. 2.





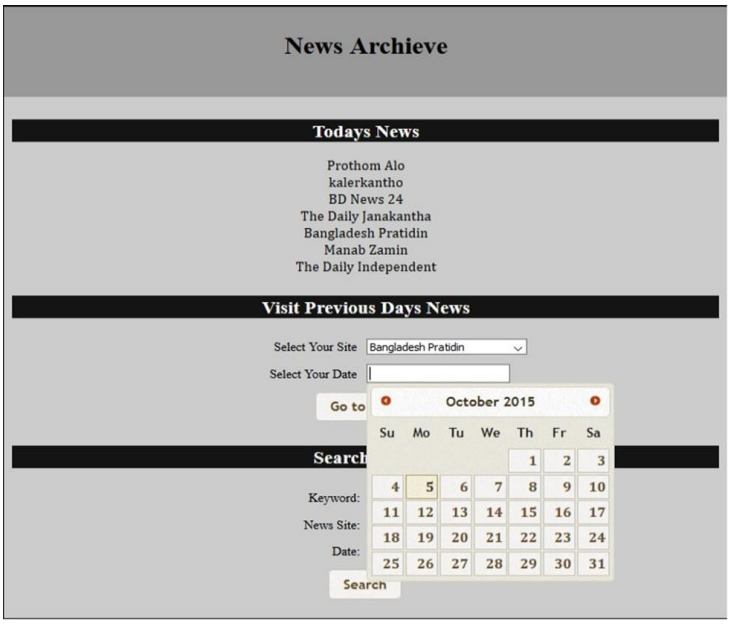


Fig.2: Screenshot of the user-interface

1.5. News Searching Option

Since it is easy to find the desired news through searching, a simple news searching option is implemented to search news for the user. Users will be able to search news based on keyword, date and also from a certain news site. This has been achieved by performing the following three steps.

First a program was developed to extract necessary keyword from a web page according to the news content of that page and save it in a database table for further using. This was done by analyzing the title, news headline and news content of that page. The program was implemented with the news crawling program. When the crawling program saves a page, it also extracts the keywords from the page.

The database table was designed to store the necessary information required for searching. The table contains the keywords, URL for that page, date of saving the page and also the website from which the page is saved. Finally the UI provides options for the user to enter his desired keyword, date and website to search the result. Based on the given information a query is made to look for that information in the database table, which returns links to the page containing the matched keyword. A demonstration of searching is shown in Fig. 3.

Fig. 3: Screenshot of the user-interface for searching option

III. Performance improvement

1.6. Bandwidth Minimization

We can see in our research work, in most cases for a given website there is a fixed number of CSS and JavaScript file. Each page from that website uses those same CSS and JavaScript file. Therefore if we only save these files once it will cover for rest of the pages from that website. This would save a large amount of

bandwidth and will significantly improve the speed of our crawling program. For example we will be able to save 1,363 KB of data each time we save a webpage from 'www.prothom-alo.com'.

1.7. Multi-threading

By definition multi-threading or multitasking is when multiple processes share common processing resources such as a CPU. Multi-threading extends the idea of multitasking into applications where we can subdivide specific operations within a single application into individual threads. Each of the threads can run in parallel. The OS divides processing time not only among different applications, but also among each thread within an application.

Java is a multi-threaded programming language which means we can develop multi-threaded program using Java. A multi-threaded program contains two or more parts that can run concurrently and each part can handle different task at the same time making optimal use of the available resources [5].

We have implemented multi-threading in our news crawling program to improve the performance of our program to save multiple pages simultaneously.

IV. Result and Discussion

4. 1 Structural Consideration for Online News Websites

We have enlisted some design consideration for an archiving friendly online newspaper site which is discussed follow.An online newspaper site should not use Ajax to retrieve or generate pages containing news. JavaScript or Ajax can be used to make the page more interactive but not to create the page. Since Ajax is much slower in executing requests, it greatly affects the speed of the crawling program and the decrease the efficiency.An online newspaper should strictly follow the standard HTML rule; otherwise it may cause error in execution and may lead to a chance of information loss.Every site should allow http compression which will improve the crawling efficiency and save network resources. The homepage or index should contain all the links for every division and subdivision of news groups so that the crawling program may find all section of news. The archive section of the news should be accessed only by a separate archive link.News from the days before should be accessed through the archive section only. News from one day should not appear in another day's news, since it greatly increases the complexity in determining the news publishing day.

Every page must contain meaningful keyword representing information about the content of the page which will help the user in searching the desired news.Number of linked CSS and JavaScript file should be minimum. Since the crawling program needs to send individual request for each file, a large number of file slows down the execution speed of the crawling program. A news containing page must contain only one h1 tag representing the news headline only. Other important information can be contained in h2 or any other tag. Every news site should use a standard Unicode font.There should a predetermined time for updating news in a day which may be one or more, so that the news archiving organization can perform their archiving operation timely. If a newspaper site follows this consideration, it is desired that the site will be crawled with much more ease and there is minimal chance of information loss which is our first priority in building a reliable news archive.

4.2 News Source for Archiving

So far in this chapter we have discussed about the structure of a news site to create an archive friendly environment. We planned to extract news from the website and store or display them as a mirror of the source website. But it is difficult to create a crawling program capable of saving all news from every news website and display them in a similar manner as the source website, since every website have their own style and structure. We therefore can propose a particular section of news source for archiving, in every online newspaper site. The news site will provide permission for the archive to look in that section where the news site will provide a copy of all the news for each day. This will make news archiving more easily and error free. The site can provide only news without any style .The archiving organization can provide a financial support to the news site for the permission to archive their news.

4.3 Authenticity of news Archive

The authenticity of a news archive is a great factor to be concerned about. Since the organization, archiving the news may have access to modify the news, there is a great chance of publishing malicious news which may cause serious problems in the upcoming future. This is also true for the reverse case, where the news publishing site can change the news and publish fake news. Both the situation will have a drastic result in the future. Therefore, there must be protocol for the news archives authenticity and the news site must contain the published news alongside with the updated news so the user may have access to both the first published news and the updated news.

V. Conclusion

In this work an automatic news archiving system has been developed, based on the principle of web crawling program. The crawling program which would search an online newspaper site and find newly published news and save the newly found page, can be called as a news crawler. The news crawling program is written in Java programming language. A simple User Interface is designed for users to browse news from the archive also. The user can search for his desired news according to the desired keyword, date or from a certain website.

While developing news crawling program difficulties have been faced to crawl every online news website due to different design structure of different website. A study has been made to point out the structural properties which are a must for online newspaper site that would make a site archiving friendly. Based on the study a proposal has been made for some structural consideration that should be kept in mind while designing an online newspaper site.

Although the objective of the study has been achieved successfully, more and more work should be done for an improved and robust news archiving system. A series of future work is proposed here.The archive should be extended to cover more and more online newspaper site to enrich its information content. More advanced programming technique can be used to boost the speed of the news crawling program. Parallel crawling can be implemented using several host machines to perform the crawling simultaneously.

An advanced theme based search or semantic searching mechanism can be implemented for user to find news more easily with more recommendations to look into.To remove the possibility of modifying or changing the news a mechanism like signature can be introduced for the news archiving system. The signature can only be generated by the news publishing website and would not be valid if someone changes the news content.

Acknowledgements

The Authors are thankful to Ministry of Science and Technology, Government of Bangladesh for the financial support.

References

[1]. Mike Burner,*Crawling towards eternity building an archive of the World Wide Web*, *Web Techniques, 2(5)*, May 1997.
[2]. Carlos Castillo and Ricardo Baeza-Yates, A new crawling model, *In Poster proceedings of the eleventh conference on World Wide Web*, Honolulu, Hawaii, USA, 2002.
[3]. Allan Heydon and Marc Najork .Mercator, A scalable, extensible web crawler, *World Wide Web Conference*, 2(4), April 1999, 219–229.
[4]. Jenny Edwards, Kevin S. Mccurley, and John A. Tomlin, An adaptive model for optimizing performance of an incremental web crawler, *In Proceedings of the Tenth Conference on World Wide Web, Hong Kong*, May 2001, 106–113.
[5]. Jean-Pierre Eckmann and Elisha Moses, *Curvature of co-links uncovers hidden thematic layers in the World Wide Web, PNAS, 99(9):5825–5829*, April 2002.