User-Guided Information Extraction Based on Webpage Layout

GRADUATE PROJECT REPORT

Submitted to the Faculty of

The School of Engineering & Computing Sciences

Texas A&M University-Corpus Christi

Corpus Christi, TX

In Partial Fulfillment of the Requirements for the Degree of Master of Science in

Computer Science

By

Chaitanya ChowdaryMaddukuri

Spring 2015

Dr. Longzhuang Li _____________________________
Committee Chairperson

Dr. DulalKar _____________________________
Committee Member
ABSTRACT

Some of our daily activities have been web mining, search and accessibility. These activities possess issues like eliminating noisy information and extracting informative content. The extraction process employs methods like automatic techniques and hand crafted rules. Automatic techniques have their focus on different Rule Based Extraction techniques, but the problem with implementing these techniques is that it increases the time complexity of the extraction process. On the other hand, extraction by using hand crafted rules is generally an effective technique which works by using string manipulation functions, but the preparation of these rules gets difficult and cumbersome for users. In this paper, we present a special approach which contains two steps that invoke each other. Initially it retrieves the information from the source code of the web page and stores them in various blocks and then by applying the Rule based Extraction algorithm based like the density classifier technique it forms various rules for that particular web page and further it retrieves the information from the web pages using these rules so created, Which makes the end user to save a lot of time and navigate through the web page easily and view the web page according to his own interest which achieves the main goal of this project.
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>iii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>vi</td>
</tr>
<tr>
<td>1. Background and Rationale</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction to Web Information Extraction</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Existing Applications</td>
<td>2</td>
</tr>
<tr>
<td>1.2.1 Automatic Extraction</td>
<td>2</td>
</tr>
<tr>
<td>1.2.1.1 Named Entity Mining</td>
<td>4</td>
</tr>
<tr>
<td>1.2.1.2 Web Based Mining</td>
<td>4</td>
</tr>
<tr>
<td>1.2.1.3 Vision Based Mining</td>
<td>5</td>
</tr>
<tr>
<td>1.2.2 Hand Crafted Rules</td>
<td>6</td>
</tr>
<tr>
<td>1.3 Proposed Solution</td>
<td>6</td>
</tr>
<tr>
<td>1.4 DOM Model</td>
<td>7</td>
</tr>
<tr>
<td>2. Narrative</td>
<td>9</td>
</tr>
<tr>
<td>2.1 Problem Statement</td>
<td>9</td>
</tr>
<tr>
<td>2.2 Motivation</td>
<td>9</td>
</tr>
<tr>
<td>2.3 Project Scope and Description</td>
<td>10</td>
</tr>
</tbody>
</table>
3. System Design........................................................................................................................11

3.1 Project Design and Architecture........................................................................................11

3.1.1 Project Architecture......................................................................................................11

3.1.2 Project Design...............................................................................................................14

3.2 System Requirements...........................................................................................................16

3.2.1 Functional Requirements..............................................................................................16

3.2.2 Software Requirements................................................................................................16

3.2.3 Hardware Requirements...............................................................................................16

4. System Implementation...........................................................................................................17

4.1 Environment......................................................................................................................17

4.1.1 My Eclipse....................................................................................................................18

4.1.2 Apache Tomcat Server.................................................................................................18

4.2 Density Based Classification Logic....................................................................................18

4.3 User Interface....................................................................................................................19

4.3.1 Merger Type of Extraction..........................................................................................20

4.3.2 News Type of Extraction.............................................................................................21

4.3.3 Comparison View........................................................................................................23

4.3.4 Image Type of Extraction............................................................................................24
4.3.5 Complete Type of Extraction.................................................................25

4.3.6 Shopping Type of Extraction.................................................................26

5. Testing and Evaluation.........................................................................................27

5.1 Software Testing..............................................................................................27

5.1.1 Unit Testing.................................................................................................27

5.1.2 System Testing.............................................................................................27

5.2 Testing on Merger Type of Extraction............................................................28

5.3 Testing on News Type Extraction.................................................................29

5.4 Comparison Based Testing between Merger Type and News Type..............31

5.5 Testing on Image Type of Extraction.............................................................33

5.6 Testing on Complete Type of Extraction.......................................................34

5.7 Testing on Shopping Type of Extraction.......................................................34

6. Conclusion and Future Work.............................................................................37

7. References and Bibliography.............................................................................38
LIST OF FIGURES

Figure 1.1 Decision Tree Algorithm Example.................................................................3
Figure 1.2 Document Object Model Example.................................................................8
Figure 3.1 System Architecture.....................................................................................13
Figure 3.2 User Interface..............................................................................................14
Figure 3.3 Data Flow.....................................................................................................15
Figure 4.1 Density Based Classification Logic...............................................................18
Figure 4.2 Home Screen...............................................................................................20
Figure 4.3 Merger type of Extraction.........................................................................21
Figure 4.4 News Type of Extraction..........................................................................22
Figure 4.5 Comparison View.......................................................................................23
Figure 4.6 Image Type of Extraction.........................................................................24
Figure 4.7 Complete Type of Extraction.................................................................25
Figure 4.8 Shopping Type of Extraction....................................................................26
Figure 5.1 Test Result on Merger Type of Extraction..................................................28
Figure 5.2 Original View of webpage in Merger type...............................................29
Figure 5.3 Test Result on News Type of Extraction.....................................................30
Figure 5.4 Test Case on Comparison Type.................................................................31
Figure 5.5 Original View of the Webpage in Image Type ....................................................32

Figure 5.6 Test Case on Image Type of Extraction...............................................................32

Figure 5.7 Original View of Complete Type of Extraction...................................................33

Figure 5.8 Test Case on Complete Type of Extraction........................................................34

Figure 5.9 Original View of Shopping Type of Extraction..................................................35

Figure 5.10 Test Case on Shopping Type of Extraction.......................................................36
1. BACKGROUND AND RATIONALE

1.1 Introduction to Web Information Extraction

The internet has been a major information source which is distributed over web pages in terms of topology as well as the diversity of contents and services. There are a lot of HTML(Hypertext Markup Language), images and other media files on the internet. The web hence serves as an invaluable source of data for studies. Hence, extracting the interesting content has become a necessity. On the other hand, web pages usually consist of noisy content, which includes advertisements, banners and unnecessary links which can usually affect the performance of systems which are based on text such as search engine, web summarization, question answering and text understanding. In this instance, the results of these techniques can be enhanced using the informative content such as text content, headline, date or author name. However, extracting informative content automatically is difficult because a web page consists of both irrelevant and informative content in the same file. This file consists of HTML tags and content in between these tags help us to display the pages in the web browsers [1].

The term data mining means drilling data, the purpose of which, is to extract an element of knowledge. Data mining is a process used to explore and analyze, by automatic or semi-automatic methods, large amounts of data. But, data mining cannot solve all problems or needs of business.

Generally, web mining has 2 main objectives. First one is to improve and develop websites, which focuses on analyzing and understanding of the user behavior on the web sites.
This can help in enhancing the content of sites by improving the organization and performance of sites.

The next one is the customization period. Here the data is collected from the websites and the data mining techniques are applied on it to extract relevant information from the sites by the users.

1.2 Existing Applications

There are a few applications currently available for filtering data from the web pages. Alongside, there are various research efforts going on to find more efficient approaches for the same. But any approach created for this web scraping has to follow one of the following two methods.

1. Automatic Extraction by forming Rules
2. Hand Crafted Methods

1.2.1 Automatic Extraction

Automatic Extraction, as the name suggests, extracts data automatically from websites. The main focus of the automatic extraction is inference through features extracted from HTML pages. The process of extracting information takes a procedure of initially forming some rules and then using the rules to extract relevant information. The automatic extraction uses Rule based Extraction techniques for extracting the required information.

Rule Based Extraction Methods
Rule based Extraction Methods have been developed to alleviate these problems by acquiring knowledge automatically from the source data. Rule based Extraction is any process that the system improves its performance.

The Rule based Extraction methods uses a few efficient Techniques like the Density classification on a dataset which is perfectly collected. A dataset is a collection of data from HTML version of a particular web page. In most cases, the DOM model or the HTML model of a particular webpage is used as a dataset.

**Decision Tree Algorithm**

A decision tree is a predictive Rule Based Extraction model that can be used to classify the required information from the data. Each internal node is considered as attribute and branches between the nodes are possible values, for example, the Quinlan ID3(iterative Dichotomiser3) decision tree building algorithm produces a decision tree that attempts to classify all the objects correctly from a given set of objects. At each step, the algorithm finds the attribute that best divides the objects into different classes by minimizing entropy after all the attributes have been classified, the results can be represented by a decision tree or a set of production rules.
Once the Decision tree or the set of production rules are formed, the required informative content is then extracted from the web pages.

Some of the approaches which uses automatic extraction techniques are mentioned below:

1. Named Entity mining
2. Web Based Segmentation
3. Vision Based Segmentation

1.2.1.1 Named Entity Mining

Named Entity Mining (NEM) is a task based text mining in which the information on the named entities of a class is mined from large amounts of data. Some examples for classes can be movie, game, book and music. Examples of classes can be to mine all the titles of movies in a text based data collection. NEM is extremely useful in many applications such as web search, online advertisement, and recommender system. For instance, we can identify the named entities in search queries using the results of NEM, and then it is likely that we can understand the user's interests and internet trends better, and deliver search results and ads which can be more relevant and interesting to users. Although there have been some studies on NEM and some progresses were made, some further investigations on the problems are still required [3].

1.2.1.2 Web Based Segmentation

The essential steps in structured web data extraction are web page segmentation and data cleaning. Identifying the main content region of a web page, removing the less important
(menus, ads, etc.), can improve the performance of extraction process by a very large extent. This is a novel and fully automatic algorithm that uses a Tag Path Sequence (TPS) representation of the web page.

Tag Path Sequence (TPS): A TPS is a string which describes the absolute path from the root of the DOM tree to a given particular node.

The TPS normally consists of a sequence of symbols (string), where each one represents a different tag path. The proposed technique here works by searching for positions in the TPS where it is possible to split in two regions where each region's alphabet do not intersect, which means that they contain entirely different sets of tag paths and thus are different regions altogether [4].

1.2.1.3 Vision Based Segmentation

Vision based segmentation is a hierarchal architecture which uses VIPS (Vision Based Segmentation) algorithm for extracting the information. VIPS is an algorithm used to extract semantic structure for a web page. This semantic structure is a hierarchal structure where each node corresponds to a block. Each node will have a value assigned (Degree of Coherence) to indicate how coherent is the content in the block based on visual perception. The VIPS algorithm makes complete use of the page layout feature. First it extracts all suitable blocks from HTML DOM tree and then tries to find out separators between these extracted blocks. Here, separators denote horizontal or vertical lines in a web page which visually crosses with no blocks. Finally, on the basis of the separators, the semantic structure for the webpage is constructed. VIPS algorithm works by employing a top down approach [5].

Advantages:
Formation of rules using Rule Extraction methods is more accurate.

Disadvantages:

1. It requires regular expressions, position constraints.
2. The time required for processing each query is higher.
3. The space required for storing information is higher.

1.2.2 Hand Crafted Rules

Hand Crafted Rules [6] is the second data mining approach for extracting relevant information from the web pages. This follows the same procedure as that of the first method which includes creating rules and collecting the informative content by applying these rules and displaying them in the end user interface. But, the difference here is ML methods are not included in the rule formation phase. This method uses specialized functions called the string manipulation functions. Some of the string manipulation functions include:

- Strconcat : combines 2 strings
- Strcpy : copies string from one place to another
- strlen : displays length of the string
- strcmp : compares two strings etc.

By making use of these string manipulation functions, it forms the rules for a particular website. And then by using the same string manipulation functions again, it retrieves the informative content by avoiding the noisy content and then displays it on the end user screen.

Advantages:

1. The time and space consumption is less when compared to the automatic extraction approach.
Disadvantages:

1. Forming rules and extracting the informative content by applying these rules and using string manipulation functions are highly complicated.
2. Forming rules using hand crafted techniques is not so accurate as in ML methods.

1.3 Proposed Solution

As seen in the existing approaches above, in order to extract relevant information and most particularly retrieving the defined search information, one of the two approaches i.e., automatic extraction or hand crafted methods have to be used. Using these techniques individually either consumes a lot of space and time or makes the procedure very complicated. So, here is an approach which uses both the techniques i.e., the automatic extraction as well as the hand crafted methods which consume less time and space and increases accuracy and in the same way decreases the level of complicacy which is our main goal.

1.4 DOM (Document Object Model)

DOM is a standard for creating and manipulating in-memory representations of HTML (and XML) content. A DOM model is created by parsing a website's source code i.e., HTML code into a HTML parser. Here, in our approach, we use either a DOM model or a HTML model as our dataset. We then apply the ML methods on these to get rules for a particular website. In real time, by parsing a HTML code into a HTML parser, it not only extracts information from larger logical units but also manipulates smaller units such as links, etc.
This model is highly editable and can easily be used to reconstruct the website.

(a) An HTML View  (b) A DOM View  (c) A Browser View

Figure 1.2 Document Object Model Example
2. NARRATIVE

2.1 Problem Statement

It is clearly seen from the previous observations that, all the current approaches are using one of the two techniques to extract the relevant information from the web sites, but these websites are running for quite a long time, but the end user is not completely satisfied as there are a vast number of user requirements in this phenomenon. Also, the designer is not completely up to his/her satisfaction as either the approach is completely complicated or it consumes a lot of space or time. This new approach which satisfies both end user and the web scrapper by using both techniques, first one for rules which save time and hand crafted techniques for reducing the complexity of extraction [1].

2.2 Motivation

In general, a web user faces a lot many problems while using the web pages where he/she wants to retrieve his/her defined search information. He/she has to face all the noisy information which is not at all linked to his defined search and the existing applications worked for many years with a desire to satisfy the user but these techniques could not satisfy the user as these approaches are basically confined to a particular section like the named entity section or the location based or the web based or the vision based and which is highly time and space consuming, and on the other hand, the phenomena is complicated with the hand crafted rules. So, this approach is developed with a motivation that the developer of this approach is completely satisfied with the space and time, And the end user is satisfied with his defined search i.e., he/she is able to search for information efficiently and avoids any noisy information and see the content what he/she exactly wants to see on his/her screen [7][11].
2.3 Project Scope and Description

From the previous observations, it is seen that forming the rules based on one particular section is not sufficient for the extraction process like the named or web based or the extraction based on the end user interest. So in this approach the rules are formed for a particular website by initially considering the HTML structure or the DOM structure of that website and perhaps the most interesting point here is that in the DOM model the DIV tags are considered as blocks and all the relevant and the irrelevant informative tags are shown under this. And on this, the rules are formed and the informative content of the website is retrieved. So, by using this approach the end user can avoid the links, advertisements and the other irrelevant information and use the website completely for his defined search. And as the web designer for a particular website makes changes, like the version upgrades but mostly the web structure would be the same. Once some rules are formed for a particular website it works for mostly all the versions and upgrades of that website [10][11].
3. System Design

Here, the system design explains about the architectural flow and the steps taken to implement the main task.

3.1 Project Design and Architecture

3.1.1 Project Architecture

Initially the main intention of this architecture is to extract the informative content avoiding the noisy or unrelated content. Avoiding the noisy content is easy for one particular website but implementing this on multiple websites makes the task harder. Therefore, Some rules have to be formed to execute on all the websites and bring the informative content. For instance, let us categorize the websites as shopping websites, news websites, sports websites, etc. The rules formed for any one of these categories should be executable on all the websites falling under that category. This means, the rules formed for www.amazon.com (shopping website) should be able to work for other websites of that category, like www.ebay.com or www.flipkart.com etc.

In Figure 3.1, There are two important steps to be followed to fulfill the task, they are as follows:


2. Extracting information using these rules

Primarily a dataset has to be formed to implement the Rule Based Extraction. This data set is formed by taking the source code of a particular website, which is generally in HTML (hypertext markup language) format and sending it to a HTML parser resulting in document object model. A document object model is application programming interface (API) which defines the logical structure of the code and the way the document is accessed and manipulated.
Once this document object model is formed the Rule based Extraction techniques has to be applied on this model. Prior to applying the Rule based Extraction techniques it checks whether there are any pre-formed rules for this website. If there are any pre-formed rules it checks whether the rules formed are applicable for all the other websites and if not then the Rule based Extraction techniques are implemented. [12].

The Density based classification basically considers the content in the web page into three different blocks i.e. the previous block, the current block and the next block. The classifier first checks the density of the links by using various values. If the density of the link is greater than that particular value then it considers the link as uninformative content, if the density of the link is less than that particular value then it checks with other values, and also the text density of the other blocks with some particular values and then step two gets invoked to create a Well Formed Document which is a simple XML document or in other words can be called a set of rules. Now these rules created can be implemented on all the websites of that category and finally by using this xml document the informative content is extracted and displayed to the end user[9] [12].
Figure 3.1: System Architecture
3.1.2 Project Design

A Project design is data flowchart of how the user interface works. In this project an efficient user interface is created where the user can easily retrieve the information required. Firstly the user have to confirm whether the URL he/she is looking for is a valid one and then the URL is pasted in the search box in this approach called web based Mining. Once the URL is pasted then he/she has to select the type in which they want to view the page as the approach consists of six types Merger, News, Compare, Images, Complete, Shopping and then press the extract button arranged at the bottom of the drop down box and can view the web page as he/she desired. Figure 3.2 gives a simple view of the user interface with the drop down box and Figure 3.3 shows the data flow [13].

![Web Data Mining](image)

**Figure 3.2: User Interface**
Figure 3.3: Data Flow
3.2 System Requirements

3.2.1 Functional Requirements

- The URL selected must be a valid one.
- The end user have to categorize his interest in viewing the web page.
- The output expected must not contain advertisements.

3.2.2 Software Requirements

- Language : J2EE
- Version : JDK 1.8
- IDE : My Eclipse
- Framework : Struts 1.2
- Server : Tomcat

3.2.3 Hardware Requirements

- Processor : PENTIUM IV
- Ram Capacity : MIN 512 MB
- Hard Disk Drive : MIN 250 GN
4. System Implementation

From the architectural diagram as shown by Figure 3.1, the dataset is formed from the source code which is the Document Object Model (DOM model) and checked whether the rules are already formed and are appropriate or not. If not then the Density based classification techniques are implemented to form the rules for that particular website. It works with the density classifier technique which checks the density of all the text content and the links content and compares that with some fixed text and link density values and forms an XML document.

In the view of the user interface which starts with the title Web Mining Application, contains a search bar where the user can copy the path of his interest and paste it here and select the category in which he wants to view the web page. Currently there are six categories like Merger, News, compare, images, complete, shopping. After selecting one of these categories there is an extract button arranged at the bottom which extracts the information according to the user interest and displays it on the screen.

4.1 Environment

The reporting system is implemented using J2EE with version JDK 1.8 in My eclipse IDE environment as they are more suitable for programming. Struts framework is used for the user interface of the desktop application. Struts framework is a standard for developing Well Architected Web applications. Apache Tomcat server is used here to open the user interface in the local host. Here the user can just browse his required URL in any browser and paste in the application and see the results.
4.1.1 My Eclipse

My Eclipse is an integrated development environment (IDE) and is commercially available on Java EE and Ajax IDE. My eclipse is built on Eclipse platform which combines both proprietary and open source code into the development environment. My Eclipse is mainly used for java and also used with other languages like C, C++ etc., and is basically developed in java and can run on Windows, Mac, Linux/GTK [16].

4.1.2 Apache Tomcat Server

Apache Tomcat is an open source web server developed in java and is inbuilt in My Eclipse which Tomcat implements various Java EE applications like JSP (Java Server pages), Java EL, Web Socket, Java Servlet and gives a pure java HTTP web server environment for Java code to run in [17].

4.2 Density Based Classification Logic

```java
if (curr.getTextDensity() <= 0.333333) {
    if (prev.getTextDensity() <= 0.555556) {
        if (next.getTextDensity() <= 9) {
            if (prev.getTextDensity() <= 4) {
                isContent = false;
            } else {
                isContent = true;
            }
        } else {
            isContent = true;
        }
    } else {
        isContent = true;
    }
} else {
    if (next.getTextDensity() == 0) {
        isContent = false;
    } else {
        isContent = true;
    }
}
else {
    if (next.getTextDensity() <= 11) {
        isContent = false;
    } else {
        isContent = true;
    }
} else {
    isContent = false;
}
return curr.setIsContent(isContent);
```

Figure 4.1: Density Based Classification Logic
Figure 4.1 is the logic used by the density classifier technique. Here, it first divides the information retrieved from the web pages into various blocks and arranges these blocks into a list. The information in each block is again divided into various blocks and the first block whose index is zero is considered as the previous and the very next block is considered as the current block and the block below it is considered as the next block. This Technique first checks the link density of the current block and if it is less than 0.3 it enters into the loop and if not it returns false which means the link is uninformative. Once it enters into the loop it then checks the link density of the previous block if it is less than 0.5 then it enters into the loop else it checks the text density of the next block if it is less than 11 it declares that the content is uninformative. In this way it checks back and forth all the information in the block and retrieves only relevant information from the block and this approach is used on all the blocks and the final documental view is displayed to the end user.

4.3 User Interface

The main task this interface performs is to check the URL pasted in the search box whether it is working or not. And then the categories created in the drop down box, Merger, News, compare, images, complete and shopping one of them is selected and the relevant informative content is displayed on the main screen.
In Figure 4.1, there are several different types mentioned in the drop box which performs its respective functions. All these functions are explained below.

4.3.1 Merger Type of Extraction

Merger type of extraction is one type where it brings all the raw data from the website avoiding the noisy content like the hyperlinks, advertisements, banner, and images etc., this merger type basically brings all the large content in the web page. This is used mainly to search for the definitions or a particular term where the user need not refer to all the irrelevant information or hyperlinks displayed on the website such as, Wikipedia. So, this saves the end user ample amount of time by just referring to the original content, as the original content always stays in the large text[13].In Figure 4.2, it clearly shows the data retrieved is in raw format but it avoids all the noisy content from the web pages.

Figure 4.1: Home screen
4.3.2 News Type of Extraction

News type of extraction mainly deals with the news kind of websites such as www.cnn.com or Google news. Here, the main task is filtering the data and bringing the raw data originally; the title and also the content behind the title is displayed clearly so that the end user will have a clear idea of what the content is about. When the news websites are searched in this approach, They give the news title and its content and the next news and its content making the user feel like he is reading a newspaper without the uninformative video's or advertisements in it. So, this type of extraction saves the end users time and also focuses on his interested location in the website.
Apart from this the News type extraction can also be used for sports websites like, www.dallascowboys.com, www.cricinfo.com, and www.nfl.com etc., in general these sports web sites have lots of uninformative content like the latest updated records of various players, about old matches and so on. When the end user wants to search about the next upcoming matches, match timings, place, the venue of the match it will be a lot of trouble for him to navigate through all the page, whereas this approach gives the direct content needed which will satisfy the end user completely [10].

Figure 4.3: News Type of Extraction
4.3.3 Comparison View

This compare type clearly shows the difference between the Merger type and the news type. In Figure 4.4, the comparison between the News type and the Merger type is shown in two separate scroll down boxes. The information in the pink color is only displayed in the Merger type and the information in the black color is common in both of them and the information the green color is only displayed in the News type.

![Figure 4.4: Comparison View](image)

Figure 4.4 is the clear comparison of the merge type and the news type of data extraction.
4.3.4 Image Type of Extraction

This type of method is created to only display the images in that particular website. In general, some of the users are only interested in viewing the images as they are more attractive and informative especially the magazines prefer images more than the texts. In these cases when the end user uses this approach and clicks on the image option in the drop down box it only gives the images avoiding the text and the noisy information [14].

Figure 4.5: Image Type of Extraction
4.3.5 Complete Type of Extraction

Here, the complete type of extraction is a generalized concept of both the news and the images i.e., if some URL is copied and pasted in the search box and complete type is selected then it completely displays all the data in the website sequentially in the order of the title, the content, the images if any and the original content so the end user will be completely satisfied with the extraction. As this approach developed is the sequential step by step implementation this is the final stage where he can see the full complete structure of the webpage avoiding the noisy content[14]. Figure 4.6 is the complete view of the website with the banner image, the title and the description.

![Figure 4.6: Complete Type of Extraction](image-url)
4.3.5 Shopping Type of Extraction

This is a special type of extraction technique only used for shopping purpose. This technique is developed in such a way that it can also display the hyperlinks in the website which is only relevant to the defined search. Once the URL is pasted in the approach it will guide us to a different page of that website where the perfect searched information is displayed along with the hyperlinks so that the user will be much comfortable in navigating through the website and selecting the design of his interest and shop accordingly [15].

![Shopping Type of Extraction](image)

**Figure 4.7: Shopping Type of Extraction**
5. Testing and Evaluation

5.1 Software Testing

The process of evaluating a software product to identify the differences between given input and expected output is called software testing. The quality of a product can be judged by testing.

5.1.1 Unit Testing

Unit testing is a type of testing where a single or groups of related units are tested. It is tested for the expected outputs for the given inputs of the unit that is implemented. In this reporting system every module will be tested under unit testing i.e. we have six types of extraction in this approach, So it checks whether each type has implemented its task correctly.

5.1.2 System Testing

System testing usually verifies that the system that is being delivered meets the scope of the product along with the specification and its purpose. In this project the system testing task would be checking whether the algorithm used is functioning correctly at each phase and the information retrieved finally is appropriate and is shown according to the user interest.

The testing part can be evaluated by trying different websites for each type and verifying the original output with the expected output.

Here the compare type would be more helpful as it clearly shows the informative difference between the Merger type and the News type. The test cases for system testing will be explained along with images in the following sections.
The test cases for this approach shows what exactly will be the output for a given URL. So, here are the test cases for each type of extraction discussed in the System Design.

5.2 Testing on Merger type of Extraction

This Merger type brings all the data in the webpage avoiding the Noisy information. "http://www.foxnews.com/politics/2015/04/27/epa-spends-84000-to-study-churches-that-preach-climate-change/", this is the input URL pasted in the interface which is searched in as www.foxnews.com so the title, banner image, hyperlinks in the website, and the advertisements are discarded and the news content with the publisher is displayed. This type of Extraction is mostly used to read the large text content in the webpage.

![Image of test result on Merger type of Extraction](image-url)

**Figure 5.1: Test result on Merger type of Extraction**
5.3 Testing on News Type of Extraction

The News type of Extraction is also only concerned with text content in the webpage but not with the hyperlinks or images or other content. This is gives the article view of the webpage. Here, for example, in order to know the next upcoming matches their timings the venue of the match and the name of the ground for any sports related website is difficult by just viewing the main page of the website. So, here a cricket related URL i.e., "http://www.espncricinfo.com/" is pasted in the search box and the News type of Extraction is selected and it gives the exact content which the end user wanted to know from the website.

![Image of a cricket webpage]

Figure 5.2: Original View of the Webpage in Merger Type
5.4 Comparison Based Testing Between News and Merger Type of Extraction

The compare type is just created to show the difference between the news type and the merger type of Extraction. The News type gives the title and the related content similar to the article type of view and the Merger type here shows entirely the raw content of the web page. In the figure 5.4 the content in the plain color is the common text in both the types and the content in the red color is only displayed in the Merger type and the content in the green color is only displayed in the News type of extraction. In Figure 5.4, www.dnaindia.com is given as the input URL and in the output firstly two scroll down boxes are created first one represents the content in the Merger type and the second box represents the content in the News type and the information below represents the difference between them.
5.5 Testing on Image Type of Extraction

This type of extraction is only considered with the images in the web page only but not with text content. And this type only gives the images which are in the main content and are which are related to the defined search avoiding the banner image, The hyper linked images, and other noisy content. In Figure 5.5, A news website on Justin Bieber is given as the input URL and the text content on the webpage is avoided and only the images are extracted. Figure 5.5 gives the original view of the webpage which contains both the content and the images related to Justin Bieber and Figure 5.6 shows the clear view of the same webpage in our approach.
Figure 5.5: Original View of the Webpage in Image Type

Figure 5.6: Test Case on Image Type of Extraction
5.6 Testing on Complete Type of Extraction

The complete type of Extraction gives the clear view of the webpage which includes the title the images related to the title and the content related to the title and the image so that the end user will be comfortable in reading about what he exactly searched for. Here a news report on NASA is given as the input URL and the output contains only the title of the News and the NASA images and the content related to the news. Figure 5.7 gives the original view of the website i.e.,"http://www.dnaindia.com/scitech/report-nasa-to-use-spy-telescopes-in-dark-energy-search-2081141" and figure 5.8 gives the complete view of the web page in the current approach.

Figure 5.7: Original View of the Webpage in Complete Type
Figure 5.8: Test Case on Complete Type of Extraction

5.7 Testing on Shopping Type of Extraction

This type of extraction works more efficiently on the shopping related websites. This type of Extraction only gives the content which is exactly related to the defined search in the website avoiding the left column which says the price selection, brand selection and the right most column which shows the other links of the same brand we searched for and the customer comments which are placed right at the bottom of each product. As the content in these shopping kind of websites are hyperlinked so when the URL is pasted in the search box it guides
the end user into a new tab which contains only the products their costs and the customer review on star ratings. Figure 5.9 gives the original view of www.ebay.com where the dell laptops are searched and Figure 5.10 gives the exact view of the same webpage in the current approach.

Figure 5.9: Original View of the Webpage in Shopping Type
Figure 5.10: Test Case on Shopping Type of Extraction
6. Conclusion and Future work

Here in this project an approach is developed which brings the informative content avoiding the noisy information, advertisements, banners etc., from the web pages using the Rule Based Extraction methods and creating rules and using hand crafted techniques to retrieve the content. This approach is different from the previous approaches as it retrieves the information based on the user interest and saves the end users time which is the main goal of any web user.

There are various ideas that can be implemented in this approach in the future to make it more productive, this can also be implemented for the websites which asks for login id and password like the social networking sites, the mailing sites by creating a login page in the user interface and navigating the user to his page which also increases the security and also avoids the unwanted advertisements.
7. References and Bibliography

1. Ma Jun; Li Tihong, "XML-based Web information extraction system design and implementation," Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on , vol.8, no., pp.551,554, 9-11 July 2010


5. Xuhong Zhang; Yanqing Zhang; Jing He; Cobia, F., "Vision-Based Web Page Block Segmentation and Informative Block Detection," Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on , vol.3, no., pp.265,269, 17-20 Nov. 2013


8. Pengzhu, Ming-sheng Zhao, "Session identification algorithm for web log Mining" Management and service science (MASS), 2010 International conference on 10.1109/CMSS.2010.5576547


