AN ALGORITHM TO ELIMINATE NOISY CONTENT FROM WEB PAGES

Dissertation submitted in partial fulfillment of the requirements for the Degree of

MASTER OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

SUKHVIR KAUR

11507554

Supervisor

INDERJIT SINGH

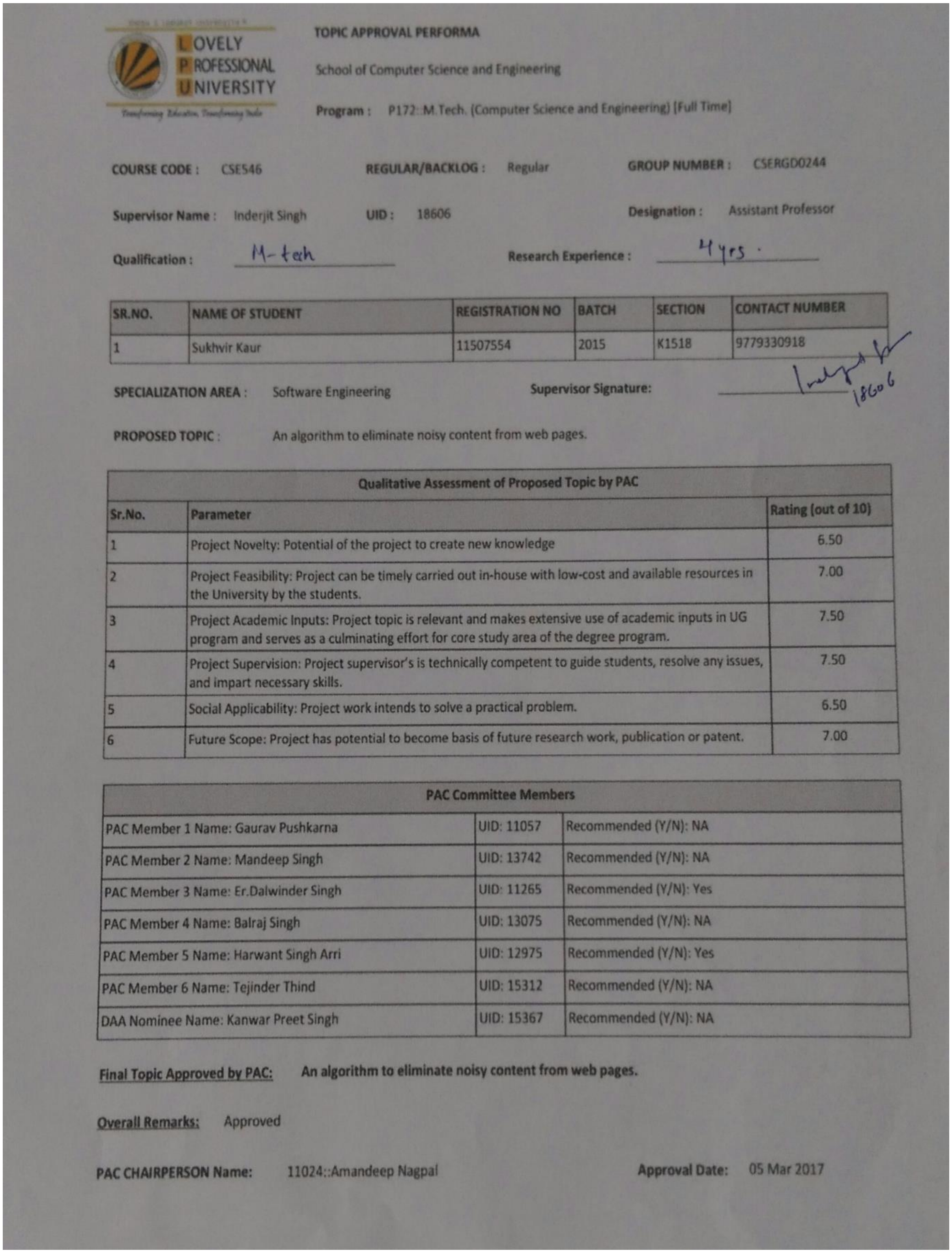
School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

May, 2017

PAC FORM



LOVELY PROFESSIONAL UNIVERSITY

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P172::M.Tech. (Computer Science and Engineering) [Full Time]

COURSE CODE : CSE546 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0244

Supervisor Name : Inderjit Singh **UID :** 18606 **Designation :** Assistant Professor

Qualification : M-tech **Research Experience :** 4 yrs.

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Sukhvir Kaur	11507554	2015	K1518	9779330918

SPECIALIZATION AREA : Software Engineering **Supervisor Signature:** 18606

PROPOSED TOPIC : An algorithm to eliminate noisy content from web pages.

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	**Parameter**	**Rating (out of 10)**
1	Project Novelty: Potential of the project to create new knowledge	6.50
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	7.00
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	7.50
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.50
5	Social Applicability: Project work intends to solve a practical problem.	6.50
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	7.00

PAC Committee Members		
PAC Member 1 Name: Gaurav Pushkarna	UID: 11057	Recommended (Y/N): NA
PAC Member 2 Name: Mandeep Singh	UID: 13742	Recommended (Y/N): NA
PAC Member 3 Name: Er.Dalwinder Singh	UID: 11265	Recommended (Y/N): Yes
PAC Member 4 Name: Balraj Singh	UID: 13075	Recommended (Y/N): NA
PAC Member 5 Name: Harwant Singh Arri	UID: 12975	Recommended (Y/N): Yes
PAC Member 6 Name: Tejinder Thind	UID: 15312	Recommended (Y/N): NA
DAA Nominee Name: Kanwar Preet Singh	UID: 15367	Recommended (Y/N): NA

Final Topic Approved by PAC: **An algorithm to eliminate noisy content from web pages.**

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11024::Amandeep Nagpal **Approval Date:** 05 Mar 2017

ABSTRACT

Data mining is the process to extract the knowledge from the database. It extracts the useful patterns from the database according to the user requirements. Thus data mining is the process to mine the meaningful data from the huge amount of data.

Web mining is the field of data mining which applies on web documents in order to achieve the particular content. Web is the large pool of data or information. Over the internet enormous amount of the database is available. This may be structured or unstructured or even be semi structured. Mining can be performed on these objects to redeem the useful content from this unorganized set of data objects. Hence this process is known as web mining. Web mining leads to the process in which meaningful data is extracted using some kind of techniques so that required content can be acquired from this vast storage of data. Web consist various thing such as web pages which involves noisy contents. Noisy contents are scroll bars, navigational bars, disclaimers like copyrights and others, advertisements. This noisy content causes the degradation of performance and reduces the efficiency of the web pages.

Purposed work consist the main part of the news from the internet newspaper. In this way it performs noise removal and scrapes the required content of specific news.

DECLARATION

I hereby declare that the research work reported in the dissertation entitled " **An Algorithm To Eliminate Noise From Web Pages**" in partial fulfillment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mr. Inderjit Singh. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

SUKHVIR KAUR

11507554

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation entitled "**An Algorithm To Eliminate Noise From Web Pages**", submitted by **Sukhvir Kaur** at **Lovely Professional University, Phagwara, India** is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

(Inderjit Singh)

Date:

Counter Signed by:

1) **Concerned HOD:**
   HoD's Signature: ________________

   HoD Name: ____________________

   Date: ___________________

2) **Neutral Examiners:**

   External Examiner

   Signature: _______________

   Name: __________________

   Affiliation: ______________

   Date: ___________________

   Internal Examiner

   Signature: _______________

   Name: __________________

   Date: ___________________

ACKNOWLEDGEMENT

I am using this opportunity to express my gratitude to everyone who supported me in this research work. First I offer my sincerest gratitude to my supervisor "**Mr. Inderjit Singh**" who has supported me throughout my dissertation .Without his dissertation would not have been written. I am very thankful for his aspiring guidelines. I invaluably constructive criticism and friendly advise during the research work. I am sincerely grateful to him for sharing their truthful and illuminating views on a number of issues related to the research. Finally, I thanks to my parents for supporting me throughout all my studies at university.

TABLE OF CONTENTS

TABLE OF CONTENTS

CONTENTS **PAGE NO.**

LIST OF FIGURES

LIST OF TABLES

TABLE NO. TABLE DISCRIPTION **PAGE NO.**

CHAPTER 1

INTRODUCTION

World Wide Web contains the enormous amount of web pages.web pages also consists some irrelevant information such as advertisements, navigations bars, copyright notice and disclaimer[1]. These are not the part of actual content thus it turns into noise. These unnecessary noises affect the quality of web pages and also decrease the performance. To provide the effective results of user's queries, it is essential to remove this noisy material and provide the particular contents required by the user. To extract the useful data by using data mining techniques from the web basically called web mining.

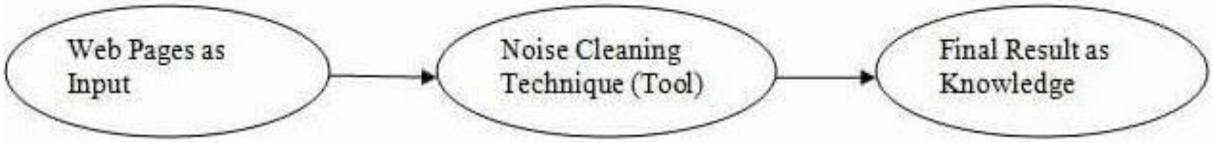


Figure 1-1: Web Mining Flow

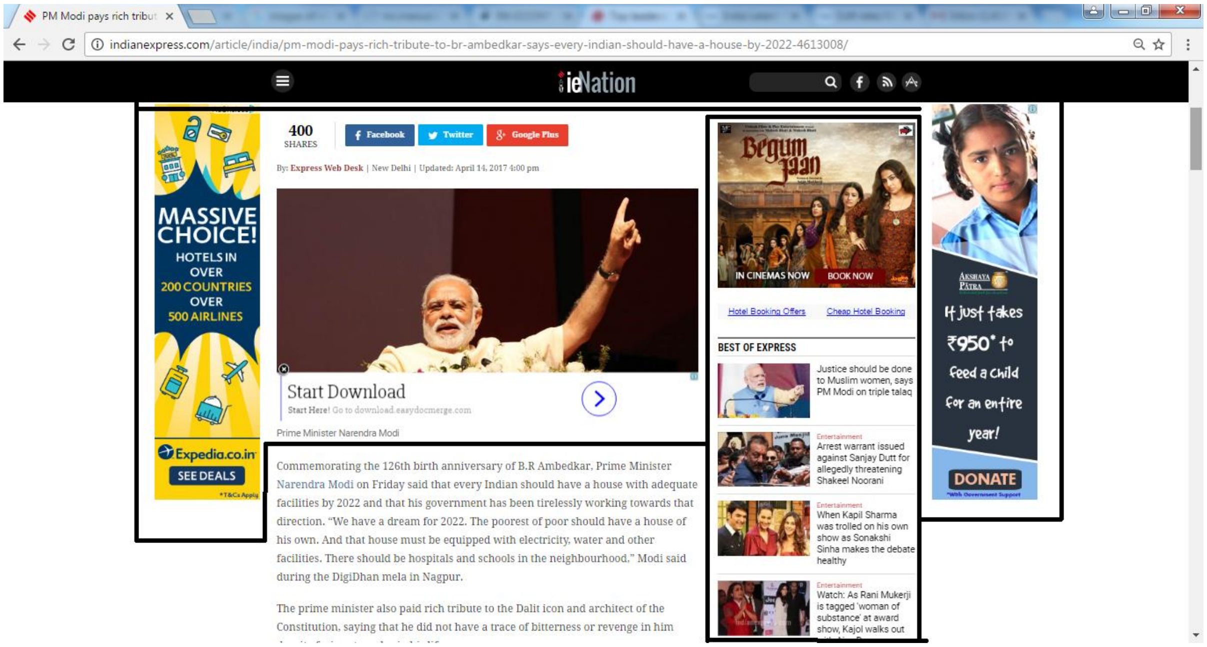
Figure 1 shows the actual processing of the web mining through the cleaning technique. First of all input is accepted where web pages are considered as input for the system. After accepting input a specific technique used to retrieve the actual part from the web pages. Thus it saves the user's time to reach at the actual information rather than searching in the entire page. The whole process helps to improve the efficiency of the page too. In this way cleaning of the web pages takes place.

Input: Web pages act as input for the cleaning process. As web pages consist extra content that causes the distraction of user from particular content.

Noise cleaning techniques: different methods are used to clean the noise of the web pages such as classification based cleaning, segmentation based, DOM based and others[2].

Final result: after cleaning the required result get produced in the form of meaningful content. In this way easily important and essential contents can be retrieved.

Example of a web page consisting noisy elements: In the image highlighted part is the noisy contents available in the web page.



By: Express Web Desk | New Delhi | Updated: April 14, 2017 4:00 pm

Prime Minister Narendra Modi

Commemorating the 126th birth anniversary of B.R Ambedkar, Prime Minister Narendra Modi on Friday said that every Indian should have a house with adequate facilities by 2022 and that his government has been tirelessly working towards that direction. "We have a dream for 2022. The poorest of poor should have a house of his own. And that house must be equipped with electricity, water and other facilities. There should be hospitals and schools in the neighbourhood," Modi said during the DigiDhan mela in Nagpur.

The prime minister also paid rich tribute to the Dalit icon and architect of the Constitution, saying that he did not have a trace of bitterness or revenge in him

Justice should be done to Muslim women, says PM Modi on triple talaq

Arrest warrant issued against Sanjay Dutt for allegedly threatening Shakeel Noorani

When Kapil Sharma was trolled on his own show as Sonakshi Sinha makes the debate healthy

Watch: As Rani Mukerji is tagged 'woman of substance' at award show, Kajol walks out

Figure 1-2: Web Page with Noisy Contents

1.1WEB MINING CATEGORIES

Web mining is classified into the three major categories. Web mining techniques are designed according to these categories. Basically the categories are used to differentiate the data available on the web pages. As web contains structured and unstructured data so these data extracted through various mechanisms, these mechanisms are basically used according to the categories of web mining. These categories are given as following[3]:

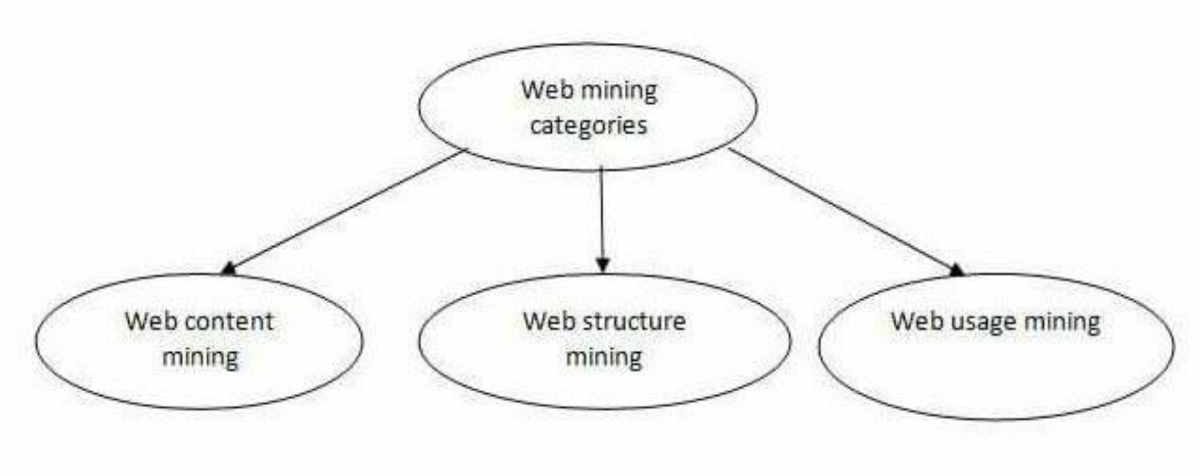


Figure 1-3: Web Mining Categories

1.1.1 Web Content Mining

It is the process in which content discovery is performed on the web pages to drive the prominent information of the page .This helps in search purposes and also executed with the web search engines[4]. Basically content mining process consists of two basic purposes such as web page content mining and search result mining.

I. **Web Page Content mining:** web page content mining consist the extraction of necessary content via extracting text, html and other multimedia. Particular approaches can be applied to retrieve the required part of the web pages rest can be discard. Because remain part contains noisy elements.

II. **Search Result Mining:** This mining process executed with the search engines, where search results are presented through the appropriate mining techniques. It shows only required results as huge amount of data available on the web associated with single entity. It is hard to provide the desired result, for this algorithm are used and present the appropriate results only.

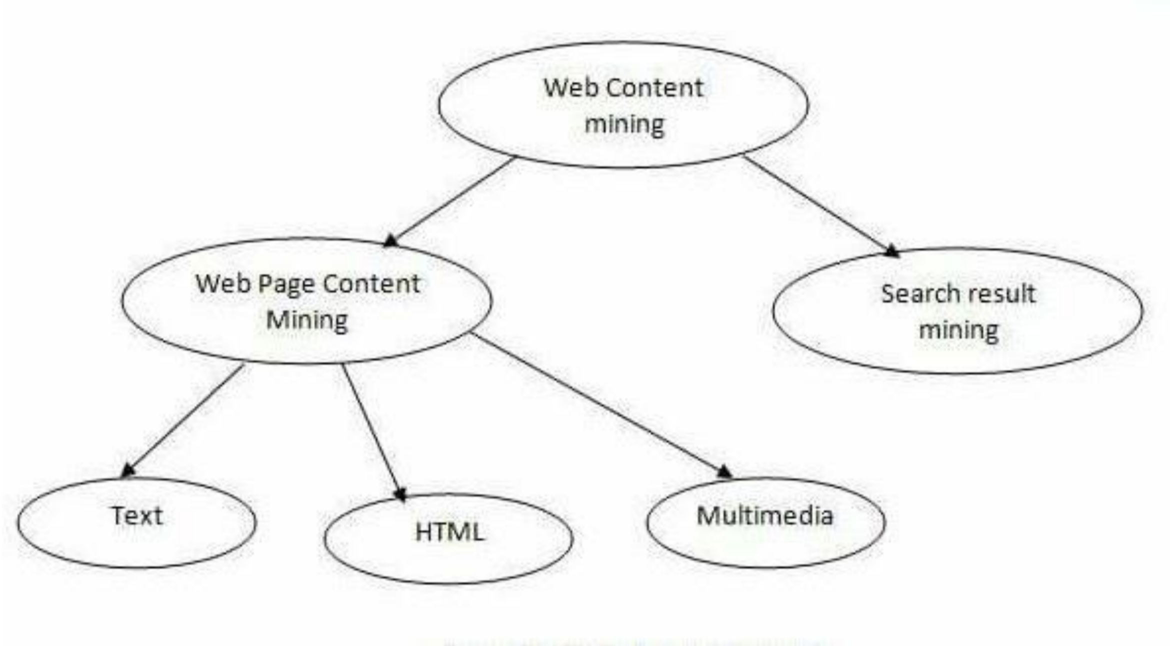


Figure 1-4: Web Content mining

Web Content Mining Approaches

Two approaches are basically used for this purpose. These are:

a) Agent based approach
b) Database approach

Both approaches functions differently from each other. Agent based approach consist further methods to detect the actual content by applying appropriate techniques. Similarly database approach deals with the noisy elements. As its name indicates it cope with the database contents such as tables and others. Database approach basically related to the database content which maintains tables and other records.

These two basic approaches are used to extract the contets.As user may require any kind of technique to retrieve .User may also need to extract the database content of particular web pages.

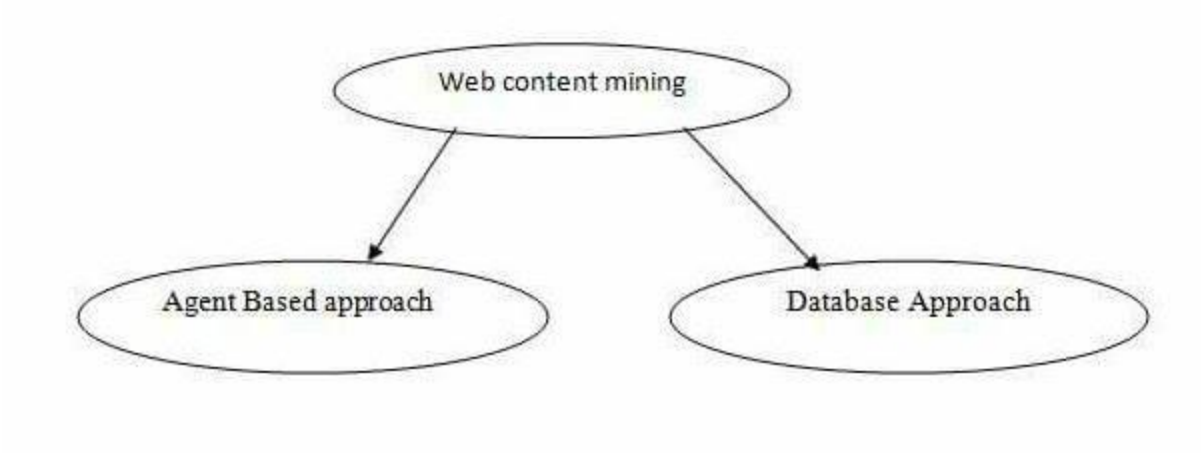


Figure 1-5: Web contents approaches

a. Agent Based Approach: This approach mainly used to extract the relevant data according to the users query .It is basically focused on the keywords that are used to represent the particular query's demand. Agent based approach involves:

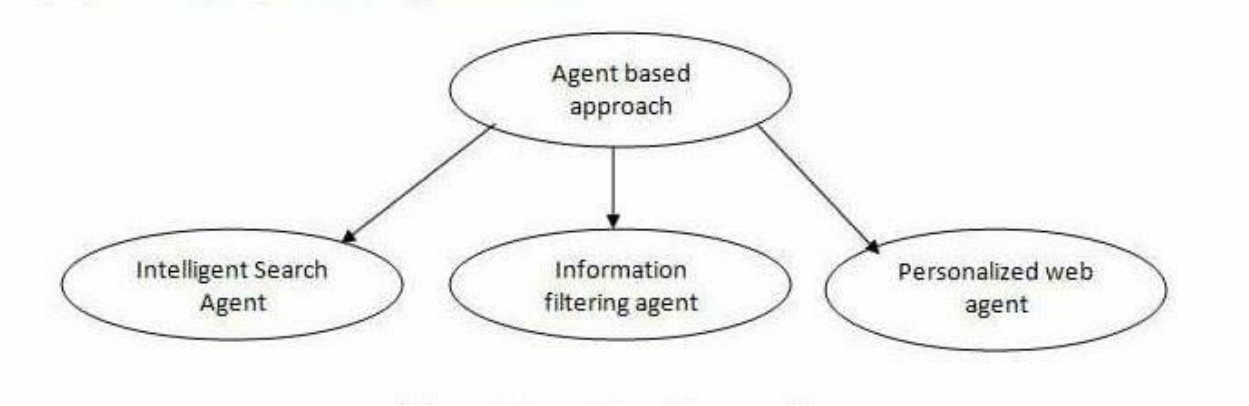


Figure 1-6: Agent Based Approach

I. **Intelligent Search Agent:** Intelligent search approach is based on the user query it is detect the query and optimize it so that result of the searching process has to be based on the query only. It optimizes the possible results intelligently.

II. **Information Filtering Agent:** This agent recognizes the elements that are essential and required. So that it filters the contents according to the informative or non informative blocks. According to their importance it generates the final output. It categorizes the data according to the usefulness of data.

III. **Personalized Web Agent:** Personalized data means it provide the documents according to the profile of the users. It detects and tracks the user activities, finally according to the result of detection documents presented to the users.

b. Database approaches

This approach related to the database consist tables, schemas, attributes with the domains. It consist a defined domain regarding these schemas, attributes and tables. It applies on the semi structured data by using techniques to organize that data and forms collection of data.

1.1.2 Web Structure Mining

This is the process of retrieving the structured information from the pages that are available on the internet. In web structure mining graph structure is preferred over the structure format.web pages are considered as the nodes and hyperlinks act as the edges thus connection among the web pages is formed. In this manner web pages get connected to each other. For this purpose algorithms are developed so that appropriate data can be derived from this immense pool of information

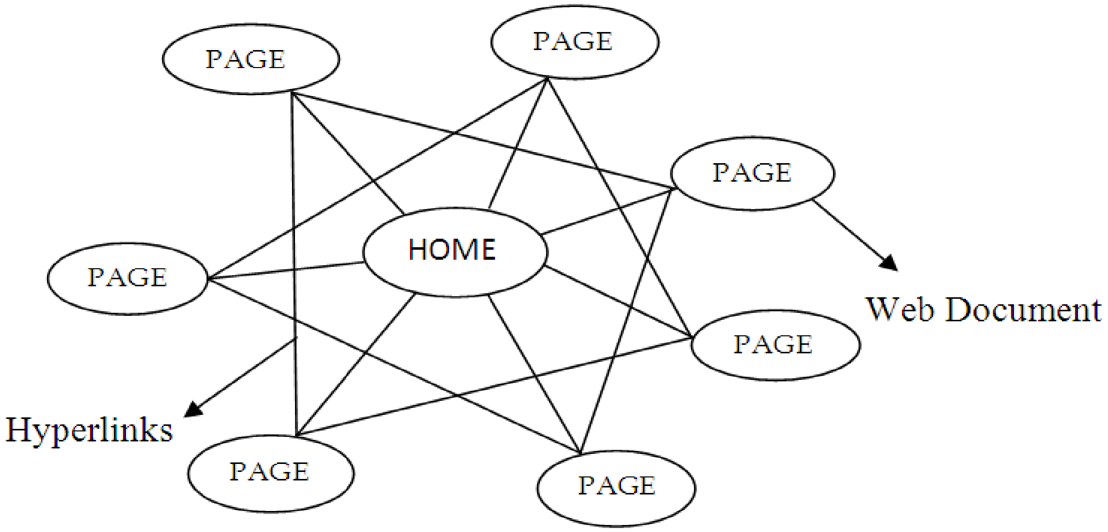


Figure 1-7: Web Structure Mining [4]

In web structure mining, it categorizes the web pages according to the different subject or contents. After categorizing it decides which web page has to contain in which collection of web pages. It can perform this process as inter-page or intra-page. Hyperlinks that indicate the different parts within the same page considered as intra-page. In Inter-page hyperlink redirect the user to new web page. Web pages can be organizes as the tree form because it becomes easy to extract different web pages. For this purpose Document Object Model (DOM) is used. Through the DOM documents can be automatically extracted. Web structure mining is used by Google, Yahoo and others. Where HITS algorithm is used by the IBM and Google uses Page Rank algorithm.

1.1.3. Web Usage Mining

It is based on the web user log mining. Where user behavior is extracted and the different patterns are developed in order to achieve the required information. This is applicable where recommendations are given to the user. In this user log are tracked according these result is provided.

Web server Logs:

I. **Access Log:** Access log is also called one of the web server logs. It focuses on the each and every activity of the user such as every click, every hit on the web pages. So it captures the essential parameters regarding user and other parameters

II. **Agent Log:** Agent logs are used to determine the behavior of user while visiting the web pages. Agent log concentrate on the user's browser version and on which operating system it is running.

III. **Error Log:** It shows error such as error 404 not found. It appears when user click on a particular link but it denied to show that link, hence shows the error.

IV. **Referrer Log:** Referrer log means it contains the URLs of other websites.

These server logs help to retrieve the required data when required. As logs are used to record the activities of the users so that data can be saved for further processing. Various organizations used these kinds of techniques to track user's behavior in order to increase their sales .These techniques basically help to grow their business. Different logs are used for extremely separate task. In this way efficiently results can be achieved.

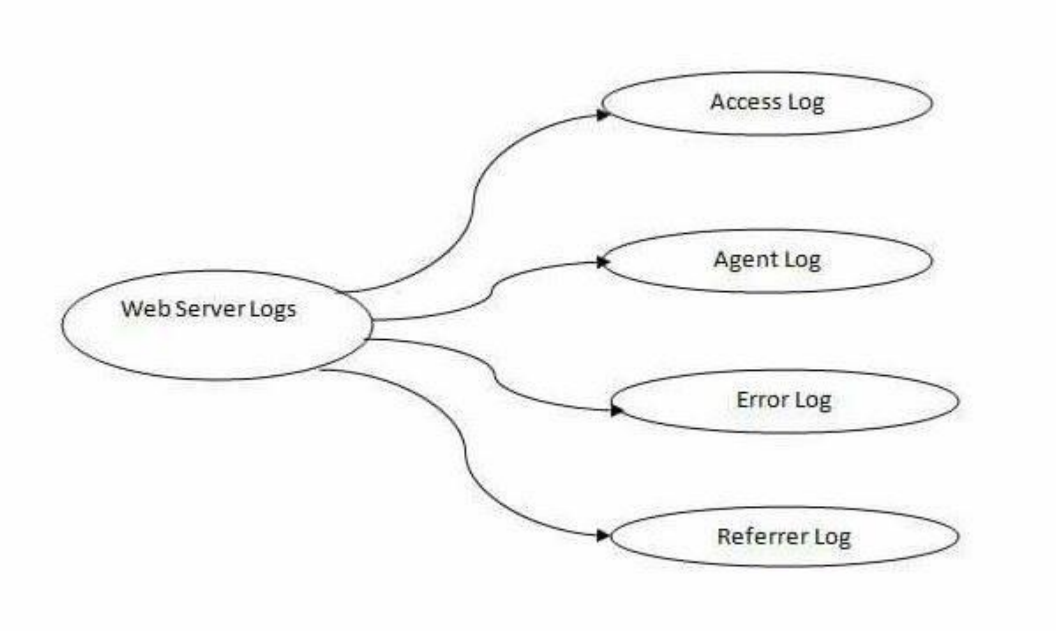


Figure 1-8: Web Server Logs

Process of the Web Usage mining

Web usage mining consists of mainly three phases. Each phase having different functionality to retrieve the required data. These phases are:

I. **Data Pre-processing:** It is the process to identify the session, users, page views and others, in order to provide a clean web page to the user. To clean the page many steps are required such as fusion, user identification by IP, authenticating data, cookies, client information.

II. **Pattern Discovery:** This is the step where data mining techniques and algorithms are used to determine the pattern .It uses algorithms such as sequential analysis, association rules, clustering, classification and others. Same pattern can be define in many ways such as graphs, charts, forms and tables

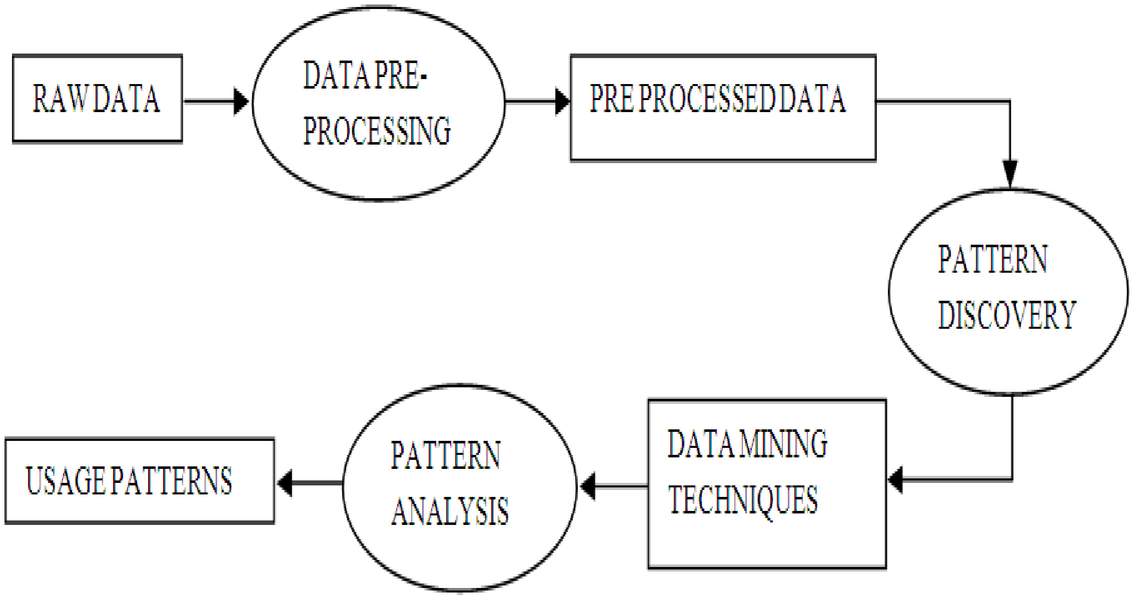


Figure 1-9: Process of web Usage mining [4]

III. Pattern Analysis: This is the final step of the web usage mining process. It uses the techniques as OLAP, visualization etc .Pattern analysis means recognizing the exact pattern by comparing with the existing one .But in the web mining process this technique is used to extract the exact pattern .Main content extracted by recognizing the important aspect of the web page and it helps to remove the noisy contents from web pages. For this purpose machine learning techniques used to identify the contents which belongs to the user's requirement. Data mining techniques are applicable for retrieving the required results as shown in the diagram. After pattern analysis exact pattern can be achieved as what kind of pattern are more trending or required by the users. These are helpful to the many businesses in order to analysis the consumer behavior over web.

1.1.4 Web Scraping (content extraction)

Web scraping is extracting the useful contents from web pages. Basically it is the process of scraping the required elements which leads to the cleaning process. users can access only the useful contents and extra noisy contents can be eliminated[5].

Web scraping process

Web scraping process involves extraction of contents via accepting the input as web pages or websites and then performs some technology in order to structure the elements of particular web page. It consist mainly few steps such as

a. First of all take web sites with html code as input.

b. To access the web pages use web scraping techniques.

c. Through web scraping technology structured data can be achieved which provide the scraped data.

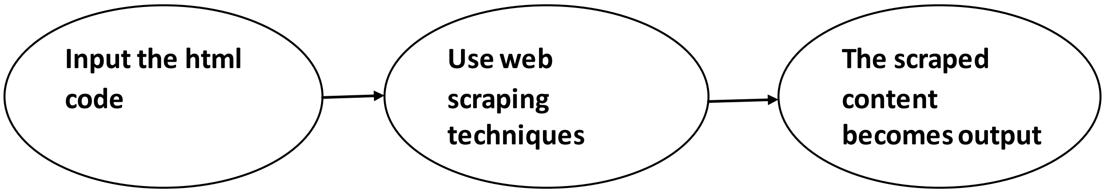

Input the html code
Use web scraping techniques
The scraped content becomes output

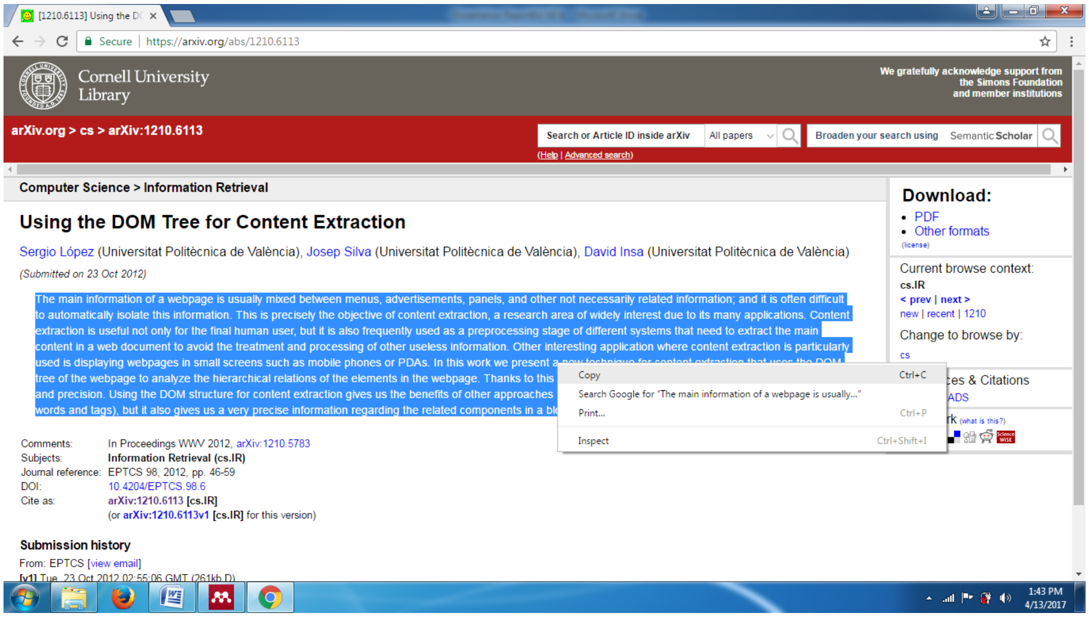
Figure 1-10 web scraping process

Web Scraping Method

Different web scraping methods are used to extract the contents from web pages in order to provide the essential contents[6].

1. **Human copy and paste**
2. **Text pattern matching**
3. **HTTP programming**
4. **HTML Parsing**
5. **DOM Parsing**
6. **Vertical aggregation**
7. **Semantic annotation recognizing**
8. **Computer Vision Web page analysis**

1. **Human copy and paste:** it is the traditional method of content scraping. Where users manually copy the content of web pages and save them for further procedure. These kind of scraping used to get particular kind of information or to generate datasets. But this method require lots of time as it is the manual process. But this is a tiresome process. Users can easily get bore from this copy and paste process and it becomes a headache at a point to do copy and then paste for large datasets.


[1210.6113] Using the DC ×
Secure | https://arxiv.org/abs/1210.6113
Cornell University Library
Search or Article ID inside arXiv
All papers
Broaden your search using Semantic Scholar
(Help | Advanced search)
Copy Ctrl+C
Search Google for "The main information of a webpage is usually..."
Print... Ctrl+P
Inspect Ctrl+Shift+I
1:43 PM
4/13/2017

We gratefully acknowledge support from
the Simons Foundation
and member institutions

arXiv.org > cs > arXiv:1210.6113

Computer Science > Information Retrieval

Using the DOM Tree for Content Extraction

Sergio López (Universitat Politècnica de València), Josep Silva (Universitat Politècnica de València), David Insa (Universitat Politècnica de València)

(Submitted on 23 Oct 2012)

The main information of a webpage is usually mixed between menus, advertisements, panels, and other not necessarily related information; and it is often difficult to automatically isolate this information. This is precisely the objective of content extraction, a research area of widely interest due to its many applications. Content extraction is useful not only for the final human user, but it is also frequently used as a preprocessing stage of different systems that need to extract the main content in a web document to avoid the treatment and processing of other useless information. Other interesting application where content extraction is particularly used is displaying webpages in small screens such as mobile phones or PDAs. In this work we present a new technique for content extraction that uses the DOM tree of the webpage to analyze the hierarchical relations of the elements in the webpage. Thanks to this
and precision. Using the DOM structure for content extraction gives us the benefits of other approaches
words and tags), but it also gives us a very precise information regarding the related components in a bl

Comments: In Proceedings WWV 2012, arXiv:1210.5783
Subjects: Information Retrieval (cs.IR)
Journal reference: EPTCS 98, 2012, pp. 46-59
DOI: 10.4204/EPTCS.98.6
Cite as: arXiv:1210.6113 [cs.IR]
(or arXiv:1210.6113v1 [cs.IR] for this version)

Submission history

From: EPTCS [view email]
[v1] Tue, 23 Oct 2012 02:55:06 GMT (261kb,D)

Download:

- PDF
- Other formats

(license)

Current browse context:
cs.IR
< prev | next >
new | recent | 1210

Change to browse by:
cs

ces & Citations
ADS

rk (what is this?)

Figure 1-11: Manual copy and paste for scraping

2. **Pattern matching procedure:** This technique is based on the regular expressions for pattern matching and then extracts the required contents. Text grouping also performed in order to recognize the main content.
3. **HTTP Programming:** This method is used to extract the main content through dynamic and static web pages. It works by posting the requests to the server using socket programming.
4. **HTML Parsing:** Data query languages and HTQL kind of languages used to extract the content from the html. First of all html parser generated then perform the extraction of elements from this parser.

Before parsing

```
<html>
<head>
<title>
“hello”
</title>
</head>
</html>
```

After parsing

```
<html>
    <head>
        <title>
             “hello”
               </title>
          </head>
    </html>
```

And other tags.

5. **DOM Parsing:** DOM is Document Object Model which provide interface to the API and browser.DOM tree used for parsing the web pages through XML on html code and then used to extract the contents required by the user.

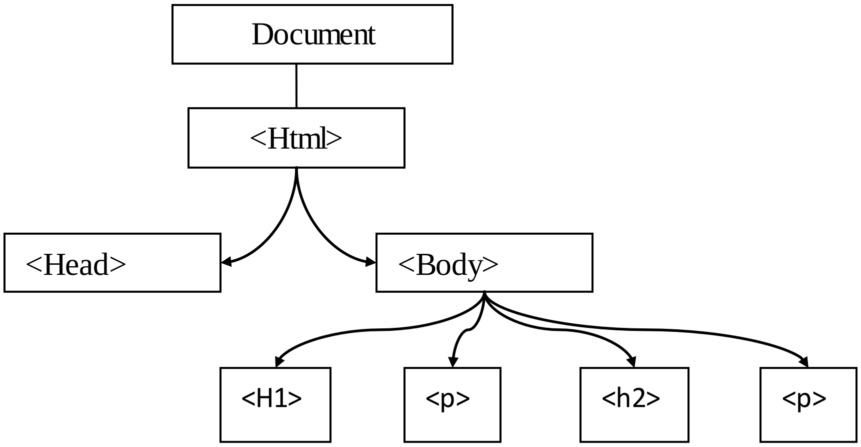


Figure 1-12: DOM Parsing

A. **DOM Tree Construction**: Source code of the web pages act as input to the DOM construction process.Html code used to construct the DOM tree from which the actual content can be determined. From the DOM tree informative and non informative contents will be detected and hence the related contents will be separated from the noisy contents.

B. **Input the DOM tree:** Dom tree will be act as input for further processing of the algorithm. Actual content can be recognized from the DOM tree only.

C. **Obtain the Content:** From the Dom tree initial nodes can be obtained. After obtaining initial nodes best nodes have to determine. Best nodes will be considered as the actual contents.

D. **Extract Sub Tree:** From the constructed DOM tree best nodes have to be calculated, these nodes will act as related content to the user query. In this way a sub tree extracted which will involve only best nodes.

E. **Final Output**: Final output will be the relevant contents which are required by the user. It will contain required text and multimedia contents only.

6. **Vertical Aggregation:** This technique is widely used by the companies having huge computing powers. This method used to target the particular vertical for scraping. These extraction methods also applicable on cloud. It creates the vertical and bots automatically therefore no human interaction required to complete the tasks.
7. **Semantic annotation recognizing:** When web pages consists the annotation then it becomes the special case of the DOM parsing. Otherwise annotation always involved with the semantic layer.
8. **Computer vision web based analysis:** It involves machine learning techniques to extract the content of web pages. It involves human interaction for interpretation of web pages.

WEB SCRAPING SOFTWARE

Various softwares are developed for scraping contents from web pages in order to remove the noise. These softwares help to retrieve the required part of web page. These softwares are automated where human intervention may require or not. Web scraping software provides the different features to extract contents from web pages, according to their features charges varies to use their facilities. Following are the web scraping software:

i. **Visual web ripper**
ii. **Web content extractor**
iii. **Mozanda web scraper**
iv. **Out wit hub**
v. **Screen scraper**
vi. **webHarvy**
vii. **Easy web extract**
viii. **webSunDew**
ix. **Helium Scraper**
x. **Scrappy**
xi. **FMiner**
xii. **Import io**
xiii. **Web scraper**

CHAPTER 2

REVIEW OF LITERATURE

Different techniques are used to clean the noise from the web pages. Noise is available in many forms in the web pages such as advertisement banners, navigational bars, copyright or disclaimer notices considered as noise. Various techniques are introduced to cope up with the noise and meaningful contents are extracted from the web pages.

G. Kaiser *et al*.[7] Define the process of extracting the essential facts through DOM tree. Their approach uses the open XML to maintain the structure of the HTML and thus form DOM tree. Content are then extracted from the formed DOM tree. For extraction purposes many filters are used .But in this paper authors describe about the two sets of filters. From which one of the set involve the Filtering of the images, scripts, styles, and links. It also avoids the some of the tag of the web pages. Second set of filters is the complex and uses the algorithmic procedure.tus it provides the high level of extraction of the contents. These filters use the advertisement remover, link list remover, empty table remover and others. After filtering the essential parts of the web page DOM tree is constructed so that appropriate content can be extracted. On this Dom tree pruning is performed according to the content after parsing the Dom white spaces are also removed by converting it into html.

D. Yang *et al*[8]. Introduced an algorithm used to extract the main content from the web page. This algorithm involves some parameters .These parameters are: Node link text density (NLTD): this is the factor which indicates the ratio of the anchor text length and all text length in a node. Non anchor text density (NATD): it is the ratio of the non link text length to total non link text length in all nodes of the page. Punctuation mark density: it is the ratio of the number of punctuation mark in a node and the punctuation mark in the whole page. These parameters are used to compute the weight of the each content block in the web page. Algorithm involves some steps such as: Arrange the page tags. Tags are arranged in such a manner that it forms the tag pair and some tags are fall under some other tags. Remove the tags that are not necessary and cause to noise in the page. Essential hyperlinks should not be

Removed .tags like <div><td><table> these needed prevention. Maintain the tag window it include all the tags which provide some prominent portion of the page. Tag window weight is computed through the parameters .this weight indicate the chances of the main content. The weight is the adjusted accordingly. It finds the weight of all tags window then chose the optimum tag window which provides the greatest weight.

H. Wang *et al* [9] For calculating the similarity of the two WebPages or HTML pages an algorithm is proposed which works on the trees. Both documents are converted into the tree format the algorithm is performed to get the results. Algorithm performed its procedure as:

First of all root nodes of the both trees are matched to each other. If the root nodes do not match or same then it depicts that both trees are not similar. If root nodes are similar then it proceeds forward.

Then sub trees are compared. First level sub tree are compared. If the value matches then save the value in the matrix. After that, the generated matrix is to be calculated. After that obtain the code of the webpage then form the DOM tree .DOM tree helps to extract data. If there are two DOM trees A and B then extraction formula may be used.

S. Lopez *et al*.[10] introduced a technique that is based on the chars nodes ratio which tells the relation between text data and tags data of each node of DOM tree .if there is a DOM tree having root node n, then Chars nodes ratio (CNR) of node n is smaller than n1.this method includes the steps such as: First of all compute the CNR of each node of the DOM tree. Choose that nodes which have the greater CNR value. Then start the traversing from those nodes which have greater CNR value. This traversing is the bottom-up. Through traversing container nodes arte searched having best ones. Because these may contain the more relevant text as possible and number of nodes may be less. These container nodes represent the HTML block. Final step is to choose the block which having more relevant content in it. Identify the main content blocks consist. To identify the main content block another algorithm is used. It takes the nodes that are identified in the first algorithm and start removing those nodes which are descendent to the others. Then it start bottom up proceeding by removing the brother nodes and it organize the parent node until its end or fix point is met. This provide a final set of nodes which consist block of the webpage.

Y. Tseng *et al* [11] developed a intelligent knowledge mining system. This system followed a approach which divided into three basic steps .Input of this method is the URL of a web page. Whereas output is the set of data or blocks that present the useful and meaningful data. The basic steps are: First main step is to convert the web page into the DOM tree and the detecting the possible informative blocks. Then features are identified of those blocks and objects are extracted. Then importance of each block is to determine and then on the basis of importance or usefulness ranking is also performed. Regular pattern of the nodes are to determine as combining the set of children nodes. After that weight of the each node group is computed. Distance between adjacent RP is also calculated. To find the meaningful blocks density is calculated. Then importance of each block is calculated. Importance is computed according to the features that are extracted. Where feature involves the values: similarity, density, diverseness.

Lan Yi *et a*l [12]proposed a technique called site style tree (SST). This is based on the layouts and actual contents of the web pages available on the web. The main idea is to construct the site style tree from the DOM tree then finally contents are extracted from that site style tree. Hence actual content is extracted through the overall developed algorithm.

S. Peak *et al*.[13] Smith introduced a classification based cleaning method. The entire cleaning method is relying upon a classifier. Classifier is based on the decision learning rules. These rules help the classifier to take the appropriate decision to differentiate between informative contents and non informative blocks of the web page.

A Panikar *et al*[14] Developed a technique that is used to perform the efficient cleaning operation on the web pages. A clean web page can be attained through the bottom up traversal of the DOM tree which contained the features only .while the bottom up traversal of the featured DOM tree weight of the nodes compared with the threshold value and elimination process done through this comparison as the weight that is less than the threshold value considered as noise and hence eviction process is performed. A parent node can be automatically considered as a noise if its children are marked as noise. In this way a web page is obtained that is free from noisy elements.

E. Silambarasan *et al*[15] Used a technique to remove the unwanted elements from the web page .To achieve this cleaning a DOM tree is constructed which is then passed through the duster framework to perform the elimination of unnecessary contents. Clustering is performed in order to differentiate the meaningful content of the web page. Clustering works on the coding to identify the hyperlinks and other elements hence clustered code will be compared with the actual code.

D. Insa *et al* [16] introduced an approach which is based on the ratio of words or leaf nodes. In this DOM tree is used as the input of the algorithm. From which actual content is retrieved. To retrieve this first of all Word Leaves ratio (WLR) computed for the each node .Then compute the relevance of the nodes. After that initial nodes and best node extracted from the tree. According to the best nodes a sub tree extracted which provide the actual related content to the user.

Madhura R. Kaddu *et al*[17] Developed a hybrid approach to extract the useful content from online web pages. Dom tree is constructed and features are extracted through this DOM tree. After feature extraction rules are generated to retrieve the informative content. Rules are generated by the automatic system and then manual rules which are infer through hand crafting uses the rules which are generated by the automatic system thus main part of the web pages are extracted. In this technique multimedia contents can also be retrieved by extraction process.

J. Alarte *et al* [18]developed a technique that is use to extract the content as template extraction. In this technique web pages are considered as the input and template is generated as an output. From the input it derives the linked WebPages that randomly or indirectly relate to each other. Thus it is the process of two phases. While the very first phase consisting the searching of the complete sub digraph in a website. Other is to extraction of template from a complete sub digraph. First step involves the searching of the complete sub digraph in the website. Basically this complete sub digraph consist the main menu of the website .It involves the all the essential links and nodes. So to determine the main menu sub digraph is derived which involves the same nodes and then nested links are also permitted in them. For this purpose authors introduces an algorithm. Second phase evolves the template extraction from the selected or determined complete sub digraph. The computation of the second algorithm is based on the two DOM tree and determine that which one consist the greater values.

Fu lei ***et al***[19] used vision based page segmentation (VIPS) method to extract the content. This method improves the content extraction by using the segment method. This method basically overcomes the shortcomings of the earlier methods such as wrapper and others. The VIPS method considered the complete layout features of the web page. Extract the appropriate blocks according to the html DOM tree structure. After that it identifies the separator for these blocks so that useful blocks can be extracted. Separators are considered as vertical or horizontal lines with these blocks. Thus extraction is purely based on these separators because semantic structure of the web page is developed through it. It separates the whole web page into independent blocks. This approach is based on top down procedure.

Gui-sheng ***et al*** [20] introduced a mechanism called template based method. This method is used to determine the theme information from the web pages. Working of this method involves Dom tree.web page is always considered as input and hence parsing has performed .Which lead to the creation of Dom trees. By using these trees contents are extracted on the basis of similarity. Similarity judging step is performed by comparing two Dom trees with each other. To do this String edit distance algorithm has used. After this page clustering has performed. Page clustering involves the common clustering techniques such as partition and hierarchical. In this way templates which are same categorized into same categories. To perform template based method the largest interaction of trees phase considered to achieve the similar templates and template trees has constructed, and then extraction performed. After template extraction, template correction takes place which involves automatic denoising and manual corrections.

Jinbeom Kang ***el al*** [21]discovered a method called repetition-based web page segmentation (REPS) by detecting tag patterns for small screen devices. This algorithm detects essential patterns in a web page and creates virtual nodes to correctly segment nested blocks. REPS work basically in for steps. The very first step involves Dom tree construction where unnecessary tags get eliminated. In the second step a sequence generated from DOM tree such as this sequence consist from a child node to root node. Third step involves the determination of the pattern from the generated sequence also recognize the candidate block by matching pattern and the sequences. In the final step the concept of virtual node has used to modify the Dom tree to achieve the desire useful contents.

Yih-Ling Hedley ***et al***[22] introduced a method known as Two-Phase Sampling (2PS).This technique is used to extract the informative blocks from dynamic documents which leads to achieving accuracy improvements. Two phase sampling as the name shows it consists two phase such as extraction and summarization of hidden databases and also validate it by prototype implementation.

Donghua Pan ***et al*** [23]proposed an approach based on link density and statistic. Approach uses some parameters such as link text density (LTD), link amount (LA), link amount density (LAD), node text length (NTL).on the basis of these parameters Content extraction involves steps such as: standardizing the web page tags, Preprocessing the web page tags, judging the location of content, finally extracting the contents and adjusting the results.

Erdinc Uzun ***et al***[24] proposed a hybrid approach for extracting informative content from web pages. This technique is based on DOM tree .This technique is based on hand crafted rules for this a model is developed .the model is based on two block tags such as DIV and TD. These tags further used for determining the informative contents from web pages. As the entire web page is based on these constructive tags, hence these are considered as the mark for determining the informative contents and generate the rules. It basically involves: Learning process, Extraction process, Rule selection and creation.

Howard J. Carey ***et al*** [25] developed a system which is called paragraph extractor that is ParEx. It works by clustering html code paragraph tags and local parent headers to find the main content of the article. It works better on the websites which used paragraph tags as this system used to scrap the main content through paragraph tags only. Because most of the websites consist useful contents under the <p> tags. It does not work over the comment section of the website. Hence to use this ParEx website must have their important contents under the <p> tags. This technology provides the better results than other methods such as Boilerpipe. They have given the future work for further improvements in the clustering algorithm.

Sandeep sirsat ***et al*** [26]purposed a method which is based on the pattern matching process. This method basically used to extract the elements of news web pages. The purposed technique of this paper uses simple heuristic for scraping the core contents from websites having semi-structured nature. This method uses the algorithm which requires regular expressions to recognize the exact pattern for content extraction process over news documents on web. It provides more accuracy in content extraction process. First of all it searches the title of the document or article by finding title tag. Then it extract the actual content through body tags as div and <p>.At last it uses algorithm in order to filter out noisy contents at achieve the core part of the article or documents. This algorithm does not depend on the DOM structure. It does not require any DOM tree to search the tags as it is highly independent over the tags. This method provides an effective way to access the news web pages and provides highly accurate results.

Kui Zhao ***et al***[27] provide an effective way to extract the blog pages on web. To achieve this goal they invented an effective blog pages extractor for better experience of the user. First of all blog page converted into the Dom tree. Through DOM tree all the necessary tags extracted such as title and body. After that a sub tree corresponding to the extracted body and title tags used to extract the features. Then SVM classifier used to eliminate the extra noisy contents.SVM applies the rules on the extracted features. At the end classifier used to find out the main part of the blog. In this way it helps to extract the content of a blog web page. As their future work they have mentioned that one can develop a method to generate a sub tree efficiently with better candidate set.

Deepak Kumar Mahto ***et al*** [28] they have shown the scraping world and explained how actually scraping takes place. They have written over the use of web scraping and developed a web crawler used to extract the main content over web pages. They have implemented a crawler which is basically consist three different phases. In the first phase web crawler will fetch the links. Second phase consist the extraction process, where extractor extract the data from these links. Third phase consist the process to store that data in csv file. Implementation for this crawler has done with python language. In their future work they have mentioned that web scraping can be used for various purposes such as price comparison websites, Big-Data Analysis.

R. Abrana ***et al*** [29] introduced a technique which is used to retrieve the web contents .It also detect the replication of the contents and filter out the contents without any noisy elements. They have introduced a hybrid algorithm in order to remove the repetitive contents of web pages. They have divided the overall work into two different portions, where first one include region separation and second one consist the information extraction process. It involves tree matching, tree alignment and other mining methods for extraction process. first of all it parse the entire document then filter out the document for contents. Then it finds tag patterns and expands those patterns. After that it performs multi string alignment which helps in region separation process. After the region separation process it classifies the extracted text. Thus the final result achieved as output of extracted text.

Suraj B.Karale *et al* [30] proposed a method which is used to extract the piece of news from web pages which consist news. Basically they have provided a method of content extraction which is a supervised method. Supervised means they used their own word bank which is used to compare the text with the existing words.Hence they perform comparison of extracted texts with those existing ones.It works on the source code of the page.Then extraction process starts .First of all it parse the html code then perform page cleaning and node filtering ,which pass through a compression process .During compression they use a word bank dictionary. After this step weight calculation of the nodes takes place.In this way they scrap the main part of the news

Table 2-1: Review of literature

Title	**Authors**	**Year**	**Approach Used**
1.DOM based content extraction of HTML documents	S. Gupta, G. Kaiser, D. Neistadt ,P. Grimm	2003	Based on two filters used to remove the extra elements such as ads. Empty tables.
2.Web content information extraction approach based on removing noise and content feature	D. Yang and J. Song	2010	Remove the unnecessary tags based on some parameters.
3.Web data extraction based on simple tree matching	H. Wang, Y. Zhang	2010	Calculate the similarity of two web pages by comparing tree of

			both web pages.
4.Using DOM tree for content extraction	S. Lopez, J. Silva and D. Insa	2012	This approach is based on the Chars Node Ratio(CNR)
5.Web mining and extraction of primary informative blocks and data objects from systematic web pages	Y. F. Tseng and H.Y. Kao	2006	Intelligent knowledge mining system
6.Eleminating noisy information in web pages for data mining	L. Yi ,B. Liu and X. Li	2003	Used Site Style Tree
7.Detecting image purpose in world wide web documents	S. Paek and J.R. Smith	1998	Classification based cleaning method
8.Noise reduction in web pages using featured DOM tree	A. Panikar, S. Panicker, P. Ravkhande, N. Tamboli, P.R .Puntambekar	2016	Bottom up traversing of DOM tree. Compare the weight with threshold
9.Removal of malicious information in web documents	E. Silambarasian, S.S. Abbirammi	2016	Used Duster framework and clustering
10.Using the words/leaves ratio in the DOM tree for content extraction	D. Insa, J. Silva and S. Tamarit	2013	Compute the word leaves ratio (WLR) ,then determine the related contents
11.Pages by using hybrid approach	M.R. Kaddu	2016	Used rules to extract text and multimedia

12.Web page template extraction based on hyperlink analysis	J. Alarte, D. Insa ,J. Silva and S. Tamarit	2015	Based on Template extraction
13.Improve the performance of the webpage content extraction using webpage segmentation algorithm	F. , M. Yao and Y. Hao	2009	Vision based page segmentation
14.A template based method for theme information extraction from web pages	G.S. Yin, G.D. Guo and J.J. Sun	2010	Template based method
15.Repititon based web page segmentation by detecting tag patterns for small screen	J. Kang,J. Yang and J. Choi	2010	Repetition based web page segmentation
16.Sampling information extraction and summarization of hidden web databases	Y.L. Hedley, A. James, M. Youns and M. Sanderson	2006	Two Phase Sampling(2PS)
17.Web page content extraction method based on link density and statistic	D. Pan, A. Qiu and D. Yin	2008	Based on some parameters. Parameters used to compute the link density of

18. A hybrid approach for extracting info. Content from web	E. Uzun, H. Volkan and T. Yerlikaya	2013	Based on hand crafted rules to detect the div and TR tags.
19.HTML Web content extraction using paragraph tags.	Howard j. Carey,III,Milos Manic	2016	Based on Paragraph tags using DOM.
20.Pattern matching for extraction of core content from web news pages	Sandeep Sirsat, Dr. Vinay Chavan	2016	Based on Pattern Matching technique.
21. Effective Blog Pages Extractor for better UGC Accessing	Kui Zhao, Yi Wang, Can Wang	2016	Based on SVM classifier.
22. A Dive into Web Scraper World	Deepak Kumar Mahto ,Lisha Soingh	2016	Implemented web crawler using python to extract text
23.A Hybrid approach for Extractin Web Information	R. abrana , S. Pradeepa		Used hybrid algorithm based on tree matcing and tree alignment
24. Extracting Brief note from internet newspaper	Suraj B. Karale , G. A. Patil	2016	Purposed algorithm based on DOM extraction

CHAPTER 3

PRESENT WORK

3.1 PROBLEM FORMULATION

Over the internet enormous amount of the database is available. This may be structured or unstructured or even be semi structured. Mining can be performed on these objects to redeem the useful content from this unorganized set of data objects. Hence this process is known as web mining. Web mining leads to the process in which meaningful data is extracted using some kind of techniques so that required content can be acquired from this vast storage of data. Web consist various thing such as web pages which involves noisy contents.

Extract the related content from the web pages to provide the efficient and great experience to the user. Clean the web pages by removing the noisy data .Eliminations of the unnecessary elements to provide the accuracy of the web pages data. Extract the textual and multimedia content for smooth accessing of the web page.

3.2 OBJECTIVES OF THE STUDY

Objectives of the study are:

Review the various techniques used to eliminate the irrelevant data and provide the actual contents.

i. To eliminate the unnecessary elements from the web pages and provide the actual content, those are related to the user query.

ii. To provide the better and efficient access to the web pages to the user.

iii. To reduce the complexity and provide the related information only.

iv. To save the time of the user while accessing web pages.

v. To improve the accuracy.

3.3 RESEARCH METHODOLOGY

Research methodology consists important phases of the system that will be used to perform the actual tasks. Each step consist the major contribution to the system in order to achieve the required contents only

PURPOSED MODEL

The proposed model has been designed for the extraction of the news text and removal of the noise data from the target page link. The proposed model works towards the removal of the unwanted components from the target webpage from the newspaper websites, and returns the plaintext form of the news, which primarily contains the news title, brief description and news body. The proposed model design can be explained with the following diagram:

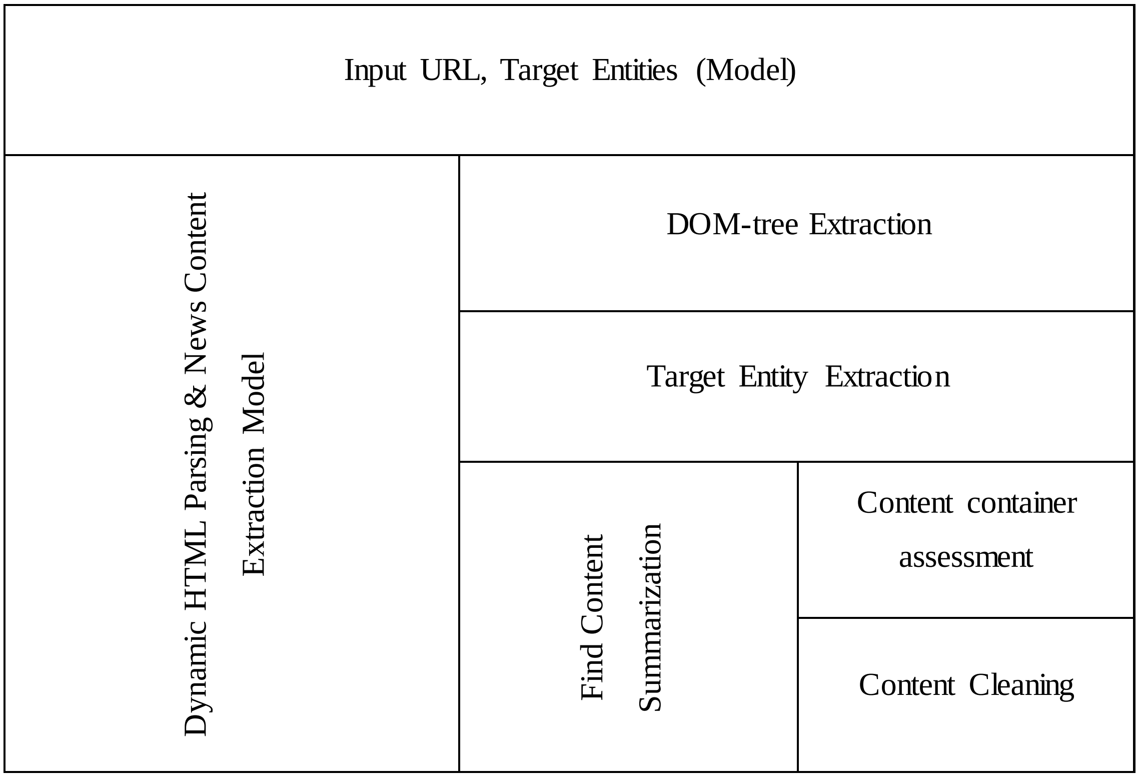

Input URL, Target Entities (Model)
Dynamic HTML Parsing & News Content Extraction Model
DOM-tree Extraction
Target Entity Extraction
Find Content Summarization
Content container assessment
Content Cleaning

Figure 3-1: Purposed Model

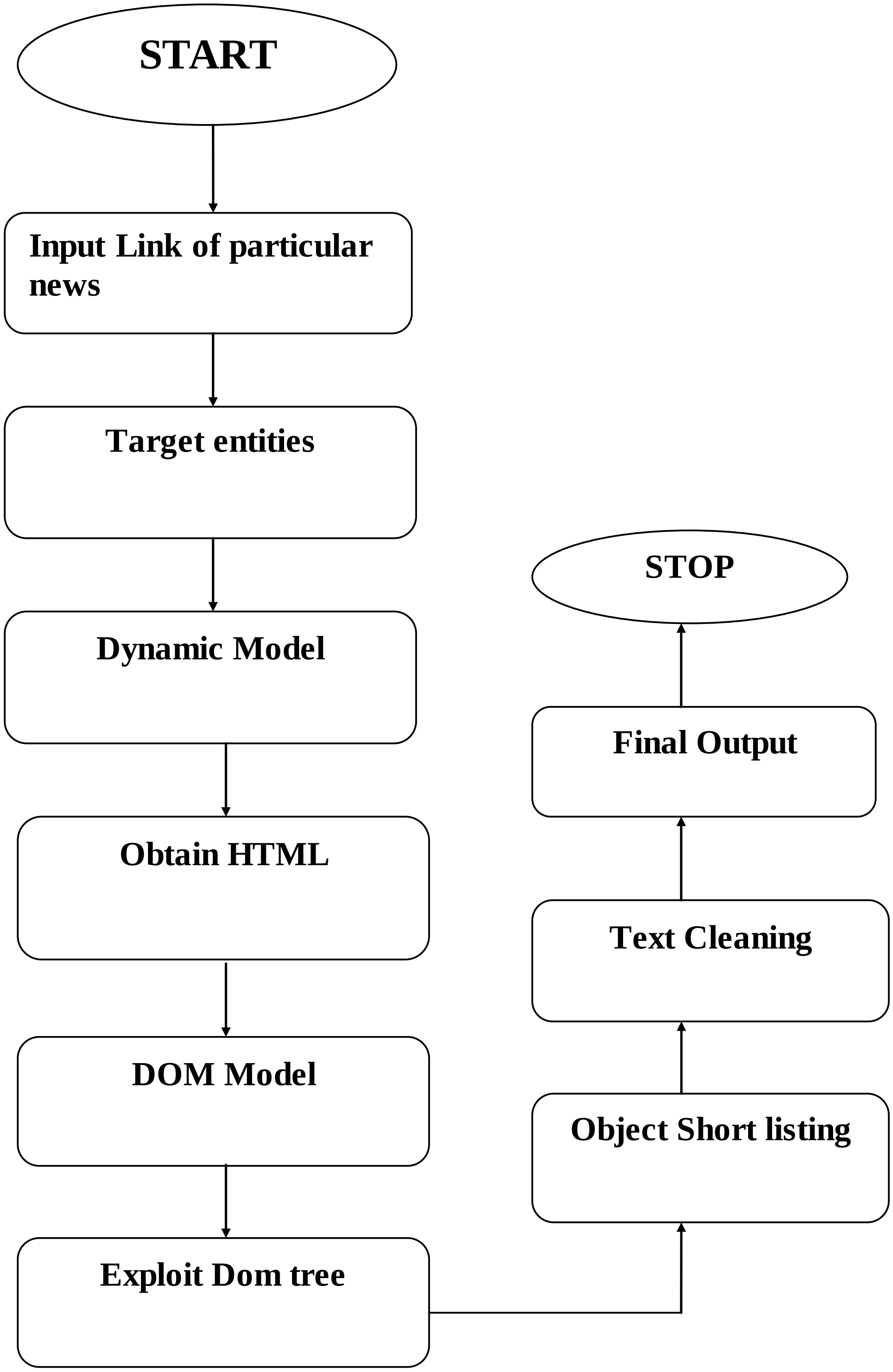


Figure 3-2: methodology of content extraction of a news

The proposed model has been designed using the multilayered formation, where the modules are designed for the acquisition of the page HTML, HTML DOM-tree parsing, DOM object selection using selective method, content extraction, noise removal, etc to achieve the goal for news content scraping from the online sources. The proposed model has been designed for the dynamic content extraction by using the dynamic and semi-supervised HTML parsing and content acquisition from the target link. The proposed model has been designed using the dynamic HTML component recognition, which utilizes the DOM-tree extraction using the "Cheerio" module under the Nodejs environment.

The DOM-tree extraction is the process of obtaining the entity tree from the HTML code obtained from the target URL, which is further parsed for the Target Entities. The target entities are extracted and cleaned using the plaintext content recognition module, which eliminates the noise (unwanted components) from the news text.

DOM-tree Extraction:

The DOM-tree extraction model works on the basis of the dependent and independent object recognition in the target HTML coding. The DOM-tree extraction model returns the HTML objects (tags) in the tree-structure, where the root always begin from the high order HTML tag of <body> or <html>, which is varied in different models. The DOM-tree model prepares the base for the content extraction model, which refers to the DOM-tree components for the recognition & extraction of the content from the target objects.

Algorithm 1: Dom Tree Extraction Model

```
1. Acquire the HTML code of the target webpage
2. Find the root tag in the HTML tree
3. Map the first level children components
4. Mark the primary tree branches
5. Iterate till the last element in the HTML element array beneath primary tree branches
    a. Recognize the one-step up (stage parent) of the target element
    b. Mark the position of element in the DOM-tree
    c. Update the DOM-tree model
6. Return the DOM-tree
```

Target Entity Extraction

The target entities are defined by the user as the input arguments for the extraction of the news contents, which involves the title, description and news body from the given URL. The user input includes the defined model for the extraction of data, which is explained with the CSS, HTML or XML entities in the target URLs. The proposed model includes the module for the extraction of the target entities defined by the user during the execution of the program. The target entities are recognized in the DOM-tree using the scraping model designed under the proposed model, which returns the entities with higher likelihood as per the input entities.

Algorithm 2: Target Entity Extraction

```
1. Acquire the target entity template → TE
2. Reform the target entity template by preparing the possible combinations
3. Acquire the DOM-tree
4. Enlist the DOM-components into array
5. Initialize the empty matching component array (MCA)
6. Iterate for the components in the DOM-tree one by one
    a. Iterate for each entity in the target entity template
        i. Match the target entity with current DOM component
        ii. If DOM component matches the target entity
            1. Update the MCA with target entity
            2. Update the MCA with matching DOM component
            3. Update the position of DOM component in MCA
        iii. Return MCA
    b. End the iteration
7. End the iteration
8. Initialize the empty Target Anchored Entity (TAE)
9. Iterate for each row in MCA → coA
    a. Iterate for each row in MCA for cross validation → coB
        i. If coA+coB == TE
            1. Update the TEA with desired combination positions
    b. End the iteration
```

```
10. End the iteration
11. Map the matching combinations in TEA
12. Extract the TEA combination based containers
13. Combine the extracted containers in the form of output array → OutArr
14. Return the OutArr
```

Content Container Assessment:

The extracted target entities are further processed using the content container assessment program, which analyzes the required content (text content in our context) for the extraction of the news components. The sub-tree is prepared from the extracted entity extracted using the previous module, which undergoes the deep analysis for the type based selection of the content carried by the extracted entities under the content container assessment.

Content Cleaning

The content is further cleaned using the content cleaning module, which primarily filters out the anchor tags, images, image captions, ad or other similar contents from the target news data. The content in the anchor tags is preserved in the body text as the plaintext, and the final plaintext is returned as the final result.

Algorithm 3: Final Content Summarization

```
1. Acquire the OutArr
2. Find the number of containers in the OutArr
3. Iterate for each container in OutArr
    a. Extract the data from each component in OutArr
    b. Mark the text data in the container
    c. Remove all remaining components (or <tags>) in container
    d. Find the text links with remaining components in container
    e. If link found with the <tags> being removed
        i. Remove the text segment connected to <tags>
    f. Add the text to clean array (CA)
4. End the iteration
5. If output target is command line
```

```
        a. Return the CA
6. Else if output target is HTML
        a. Update the output HTML file
```

SIMULATION ENVIRONMENT

The simulation model has been designed using the Nodejs model, which works upon the basis of the Javascript (Jscript) or Jquery model. The HTML code has been utilized for obtaining the output from the target URL after applying all of the results.

Table 3-1: Minimum Requirements

CPU	Intel processor with double core and 1.5 GHz of processing speed
RAM	1-4 GB
Storage (Hard Disk) Space	80-100 GB
Operating System	Microsoft Windows 7/8 or above
Web Browser Software	Google Chrome, Firefox
Programming Language/Arhcitecure/Package	Jquery, Jscript, NodeJS

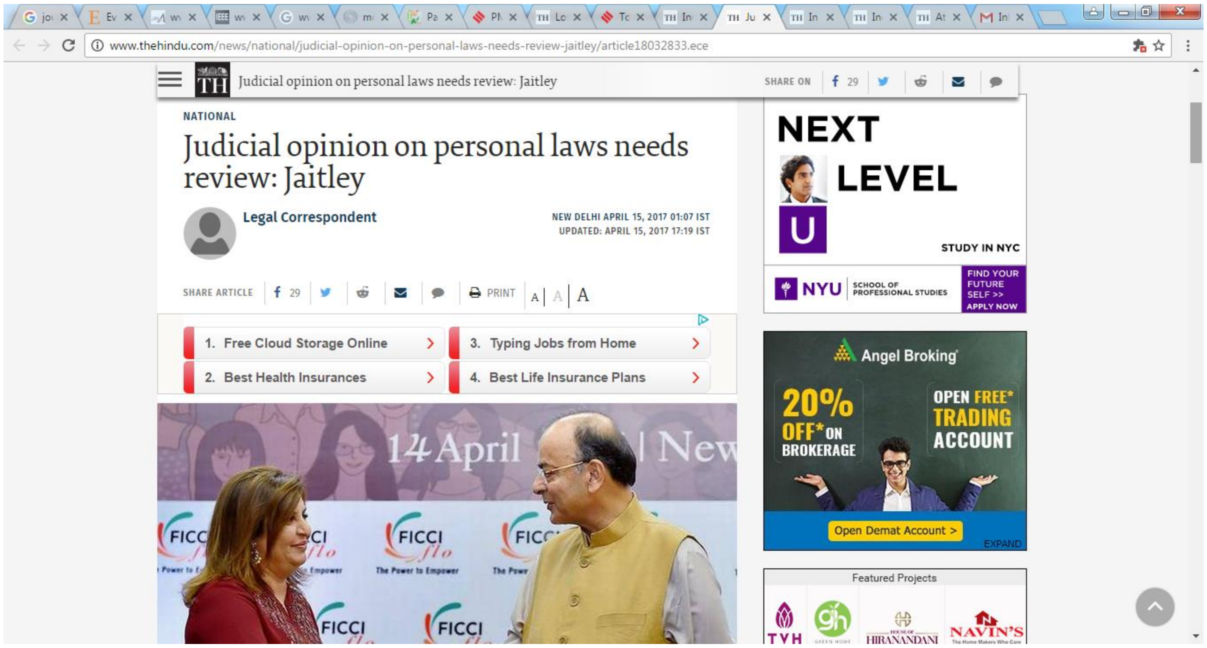
NodeJS has been used to simulate the proposed model, which has been primarily written as the module for the extraction of the content from the target URL.

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 EXPERIMENTAL RESULTS

Purposed model is basically extracts the required text only by filtering the target page. It extracts the text of that particular news of the target link. It provides that text in a html file so that user can view that specific part only without disruption of the other noisy contents. In this way it provides the required elements without any noisy contents. Experimental model consists the actual implementation of the work by showing the difference of the view of news content on actual web page and the model's working by scraping the specific part of news along with body, title and brief introduction of that particular news. It provides the effective view by extracting the piece of news.



www.thehindu.com/news/national/judicial-opinion-on-personal-laws-needs-review-jaitley/article18032833.ece

Judicial opinion on personal laws needs review: Jaitley

NATIONAL

Judicial opinion on personal laws needs review: Jaitley

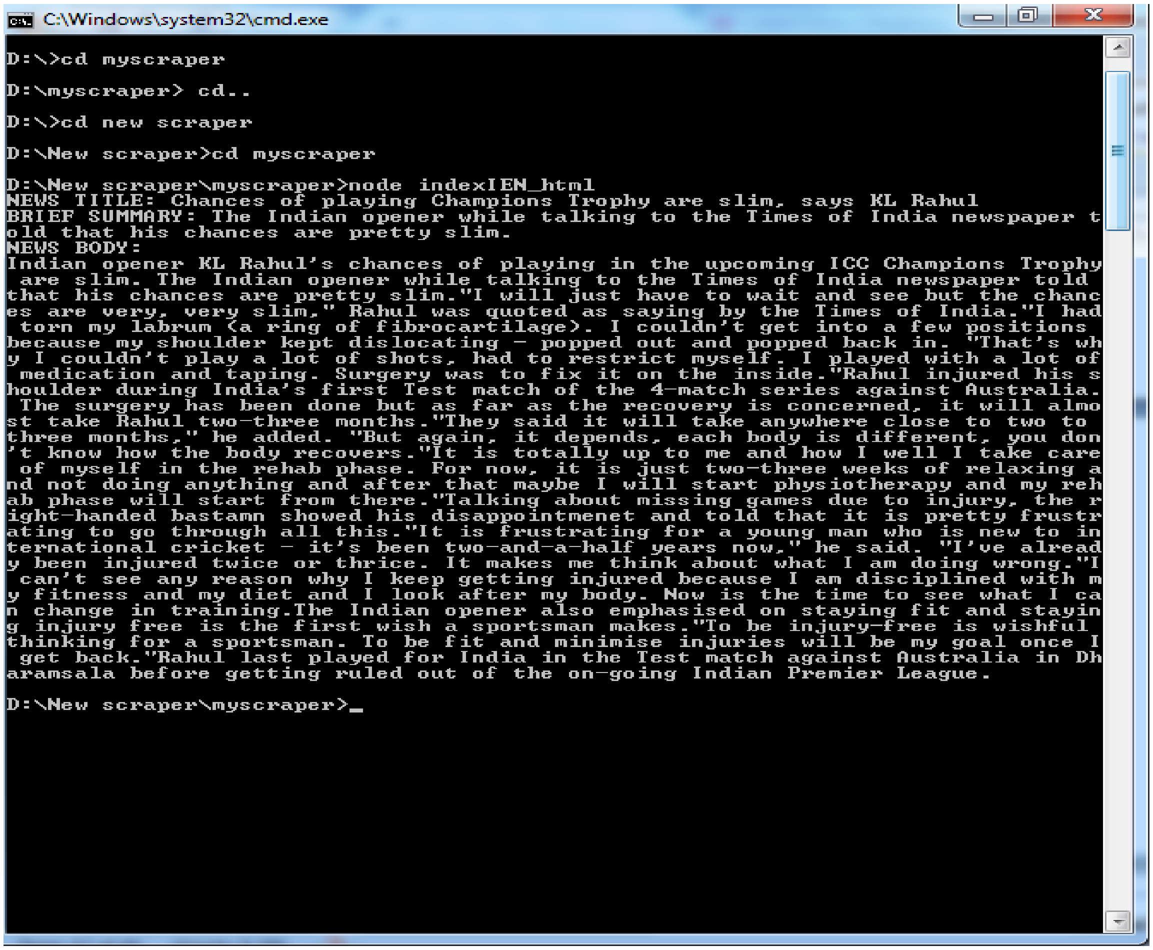
Legal Correspondent

NEW DELHI APRIL 15, 2017 01:07 IST
UPDATED: APRIL 15, 2017 17:19 IST

1. Free Cloud Storage Online
2. Best Health Insurances
3. Typing Jobs from Home
4. Best Life Insurance Plans

Figure 4-1: News with noisy contents on web page

Above image shows news with extra elements, that leads to unnecessary noise on that web page. This noise distract user to the actual contents. Therefore our model provides the required piece of news only.



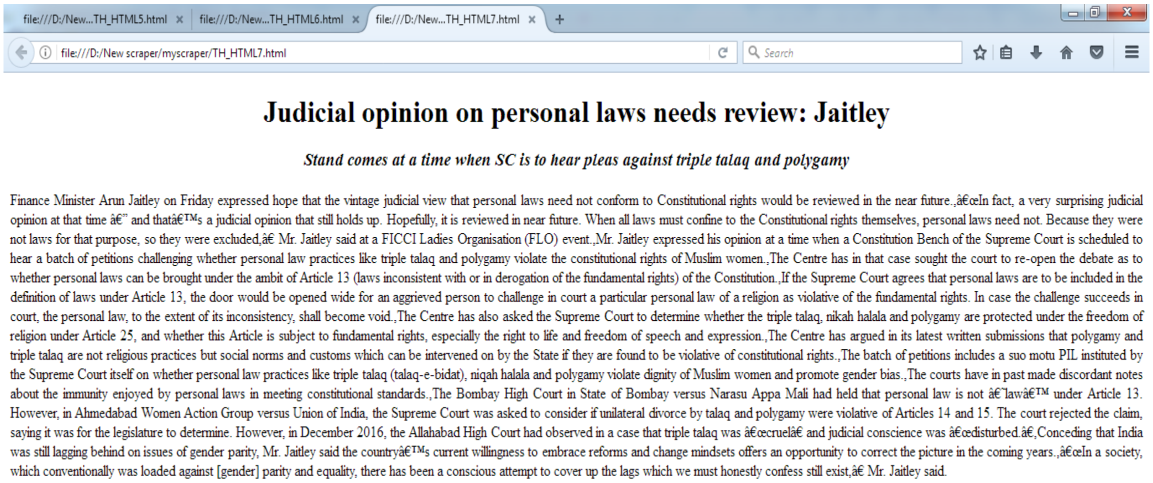
```
D:\>cd myscraper

D:\myscraper> cd..

D:\>cd new scraper

D:\New scraper>cd myscraper

D:\New scraper\myscraper>node indexIEN_html
NEWS TITLE: Chances of playing Champions Trophy are slim, says KL Rahul
BRIEF SUMMARY: The Indian opener while talking to the Times of India newspaper t
old that his chances are pretty slim.
NEWS BODY:
Indian opener KL Rahul's chances of playing in the upcoming ICC Champions Trophy
 are slim. The Indian opener while talking to the Times of India newspaper told
that his chances are pretty slim."I will just have to wait and see but the chanc
es are very, very slim," Rahul was quoted as saying by the Times of India."I had
 torn my labrum (a ring of fibrocartilage). I couldn't get into a few positions
because my shoulder kept dislocating - popped out and popped back in. "That's wh
y I couldn't play a lot of shots, had to restrict myself. I played with a lot of
 medication and taping. Surgery was to fix it on the inside."Rahul injured his s
houlder during India's first Test match of the 4-match series against Australia.
 The surgery has been done but as far as the recovery is concerned, it will almo
st take Rahul two-three months."They said it will take anywhere close to two to
three months," he added. "But again, it depends, each body is different, you don
't know how the body recovers."It is totally up to me and how I well I take care
 of myself in the rehab phase. For now, it is just two-three weeks of relaxing a
nd not doing anything and after that maybe I will start physiotherapy and my reh
ab phase will start from there."Talking about missing games due to injury, the r
ight-handed bastamn showed his disappointmenet and told that it is pretty frustr
ating to go through all this."It is frustrating for a young man who is new to in
ternational cricket - it's been two-and-a-half years now," he said. "I've alread
y been injured twice or thrice. It makes me think about what I am doing wrong."I
 can't see any reason why I keep getting injured because I am disciplined with m
y fitness and my diet and I look after my body. Now is the time to see what I ca
n change in training.The Indian opener also emphasised on staying fit and stayin
g injury free is the first wish a sportsman makes."To be injury-free is wishful
thinking for a sportsman. To be fit and minimise injuries will be my goal once I
 get back."Rahul last played for India in the Test match against Australia in Dh
aramsala before getting ruled out of the on-going Indian Premier League.

D:\New scraper\myscraper>_
```

Figure 4-2: Final Scraped Output

Command prompt used to execute the code through Nodejs. It uses nodes commands and shows the result by executing the JS file which links it to the server and retrieve the result by performing the required actions. Hence it extract the text from the target link and provide the Title of the news along with its brief introduction part. Then it retrieve the main part of news, The main part is considered to be the most required part which is always inside the main DIV container.

After scraping the contents it saves that content into an html file which provide an effective to the news text. In this way user can only see the news elements. Noisy contents such as advertisements, other news links are avoided. Hence it cleans a web page and provide the view of actual content only. It extracts the text without any further links which mostly gives into news contents. It does not leave the text with link but it retrieves that part as text form. It just avoids the link part which is also considered as the noise of web pages. Hence it cleans that particular news for the user. In this way, developed model provide a news text without noisy elements and user can experience the actual part of the desired news.

Sometime web pages consist some pop ups which cause the disturbance and distract the user, which leads to the noise for the desired part. But, my work provide an efficient way to access those required contents.


file:///D:/New...TH_HTML5.html
file:///D:/New...TH_HTML6.html
file:///D:/New...TH_HTML7.html
file:///D:/New scraper/myscraper/TH_HTML7.html
Search

Judicial opinion on personal laws needs review: Jaitley

Stand comes at a time when SC is to hear pleas against triple talaq and polygamy

Finance Minister Arun Jaitley on Friday expressed hope that the vintage judicial view that personal laws need not conform to Constitutional rights would be reviewed in the near future.,â€œIn fact, a very surprising judicial opinion at that time â€” and thatâ€™s a judicial opinion that still holds up. Hopefully, it is reviewed in near future. When all laws must confine to the Constitutional rights themselves, personal laws need not. Because they were not laws for that purpose, so they were excluded,â€ Mr. Jaitley said at a FICCI Ladies Organisation (FLO) event.,Mr. Jaitley expressed his opinion at a time when a Constitution Bench of the Supreme Court is scheduled to hear a batch of petitions challenging whether personal law practices like triple talaq and polygamy violate the constitutional rights of Muslim women.,The Centre has in that case sought the court to re-open the debate as to whether personal laws can be brought under the ambit of Article 13 (laws inconsistent with or in derogation of the fundamental rights) of the Constitution.,If the Supreme Court agrees that personal laws are to be included in the definition of laws under Article 13, the door would be opened wide for an aggrieved person to challenge in court a particular personal law of a religion as violative of the fundamental rights. In case the challenge succeeds in court, the personal law, to the extent of its inconsistency, shall become void.,The Centre has also asked the Supreme Court to determine whether the triple talaq, nikah halala and polygamy are protected under the freedom of religion under Article 25, and whether this Article is subject to fundamental rights, especially the right to life and freedom of speech and expression.,The Centre has argued in its latest written submissions that polygamy and triple talaq are not religious practices but social norms and customs which can be intervened on by the State if they are found to be violative of constitutional rights.,The batch of petitions includes a suo motu PIL instituted by the Supreme Court itself on whether personal law practices like triple talaq (talaq-e-bidat), niqah halala and polygamy violate dignity of Muslim women and promote gender bias.,The courts have in past made discordant notes about the immunity enjoyed by personal laws in meeting constitutional standards.,The Bombay High Court in State of Bombay versus Narasu Appa Mali had held that personal law is not â€˜lawâ€™ under Article 13. However, in Ahmedabad Women Action Group versus Union of India, the Supreme Court was asked to consider if unilateral divorce by talaq and polygamy were violative of Articles 14 and 15. The court rejected the claim, saying it was for the legislature to determine. However, in December 2016, the Allahabad High Court had observed in a case that triple talaq was â€œcruelâ€ and judicial conscience was â€œdisturbed.â€,Conceding that India was still lagging behind on issues of gender parity, Mr. Jaitley said the countryâ€™s current willingness to embrace reforms and change mindsets offers an opportunity to correct the picture in the coming years.,â€œIn a society, which conventionally was loaded against [gender] parity and equality, there has been a conscious attempt to cover up the lags which we must honestly confess still exist,â€ Mr. Jaitley said.

Figure 4-3: Extracted News text without noisy elements

Figure 4-3 shows that extracted text is the only text part of specific news without any noisy elements. It contains title of the news with brief introduction part and finally the body part. Body part extracted by using the <p> of the news web pages. Similarly it extracted the title by title tag of the main Div container.

The performance parameters utilized for the purpose of evaluation of the results proposed model. The statistical parameters to measure the statistical errors (Type 1 and Type 2) are measured in order to evaluate the overall performance of the proposed model by evaluating the samples by the means of the programming or the manual binary classification. The proposed model evaluation is entirely based upon this statistical analysis. The following table explains the significance of the type 1 and type 2 statistical errors for the evaluation of the hypothesis.

PERFORMANCE PARAMETERS

True Positive: The true positive is when the final condition marked as matching and correct, which shows the positive condition and denies the null hypothesis. True positive is given with the symbol A. The true positive is given as the following

$$TP = n_{11} = \text{number of such individuals}$$

True Negative: The true negative is when the final condition marked as non-matching and correct, which shows the negative condition and accepts the null hypothesis. True negative is given with the symbol B. The true positive is given as the following:

$$TN = n_{00} = \text{number of such individuals}$$

False Positive: The false positive is when the final condition marked as matching and incorrect, which shows the positive condition and denies the null hypothesis. False positive is given with the symbol C. The false positive is given as the following:

$$FP = n_{01} = \text{number of such individuals}$$

False Negative: The false negative is when the final condition marked as non-matching and incorrect, which shows the negative condition and accepts the null hypothesis. False negative is given with the symbol D. The false negative is given as the following:

$$FN = n_{01} = \text{number of such individuals}$$

Precision: The precision depicts the accuracy of the model in the presence of the false positive cases. The accuracy of the model depicts the overall impact of the false positive cases, which rejects positive cases. A positive case in our case is when the news contains the script data from one of the registered category, but returns the false result by not detecting the text in final output news text result.

I. **Precision= TP/ (TP+FP)**

Recall: Recall is the probability that a test will indicate 'test' among those with the matching sample.

II. Recall = TP/ (TP+FN)

Accuracy: The percentage of the correct result out of the Total results is called accuracy. Accuracy is also known as success rate.

III. Accuracy= (Correct Results/ Total Results) *100

F1-Measure: The F1-Measure is the cumulative parameter to assess the overall impact of the precision and recall in the case to study the overall impact of the false positive and false negative cases over the overall accuracy assessed from the preliminary statistical parameters. The F1-score value is represented in the range of 0 to 1 or 0 to 100, decided as per the maximum ranges of the precision and recall. The following equation is utilized to measure the F1-measure: **IV. F1-Measure = 2 * ((Precision * Recall) / (Precision + Recall))**

RESULT ANALYSIS

The proposed model results have been obtained in the form of various statistical errors. The word to word based analysis has been performed under the proposed model results, where the wrong results are classified under the negative results category, which is considered or classified under the micro parameters of statistical type 1 and type 2 errors. The proposed model has been also collected in the form of total words extracted, total words of news data, True positive & true negative cases and false positive & False negative cases. The following table shows the results of the proposed model obtained from the "The Hindu (TH)" newspaper scraper:

Table 4-1: Word count based analysis merged with statistical type 1 and 2 errors for 'TH'

News	Total words	Fetched words	True positive	False positive	True negative	False negative
1	299	299	299	0	1	1

2	326	326	326	0	1	0
3	401	401	401	0	1	1
4	364	364	364	0	1	1
5	573	573	573	0	1	2
6	439	439	439	0	1	1
7	125	125	125	0	1	1
8	330	330	330	0	1	0
9	218	218	218	0	1	1
10	294	294	294	0	1	1

The proposed model results have been further analyzed based upon the various accuracy parameters, which includes the Precision, Recall, F1-error and overall accuracy. The proposed model has been found efficient in the terms of proposed model results, as all of the obtained results for Precision, Recall, F1-error and overall accuracy are found above 99.65%, which posts the very high ability of the proposed model for scarping the news from the online portal of "The Hindu (TH)" newspaper. The results obtained as shown in the following table (Table :

Table 4-2: Performance analysis of the proposed model for 'TH'

News	**Precision**	**Recall**	**F1**	**Accuracy**
1	100	99.66	99.83	99.66
2	100	100	100	100
3	100	99.66	99.83	99.66
4	100	99.66	99.83	99.66

5	100	99.65	99.83	99.65
6	100	99.66	99.83	99.66
7	100	99.66	99.83	99.66
8	100	100	100	100
9	100	99.66	99.83	99.66
10	100	99.66	99.83	99.66

Results of Indian express(IE) newspaper are also considered for this purposed model. The table shows performance as with the performance parameters which gives highly associated model results. These are used to depict the performance of the model as such:

Table 4-3: Word count based analysis merged with statistical type 1 and 2 errors for 'IE'

News	**Total words**	**Fetched words**	**True positive**	**False positive**	**True negative**	**False negative**
1	557	557	557	0	1	3
2	215	215	212	3	0	0
3	490	490	490	0	1	0
4	487	485	480	4	1	2
5	225	229	218	7	4	0
6	457	457	457	0	0	1
7	420	418	418	2	0	0
8	511	508	508	3	0	1
9	504	502	502	2	0	1
10	620	618	618	2	0	1

Table 4-4: Performance analysis of the proposed model for 'Indian Express'

News	Precision	Recall	F1	Accuracy
1	100	99.46	99.73	99.46
2	98.6	100	99.2	98.6
3	100	100	100	100
4	99.18	99.59	99.38	98.78
5	96.88	100	98.4	96.94
6	100	99.78	99.89	99.78
7	99.52	100	99.76	99.52
8	99.41	99.80	99.60	99.21
9	99.60	99.80	99.69	99.41
10	99.68	99.84	99.76	99.52

4.2 COMPARISON WITHEXISTING TECHNIQUE

The performance of the proposed model has been analyzed in the form of various performance parameters, out of which the performance comparison has been evaluated using the statistical accuracy based parameters of precision, recall, F1 and overall accuracy. The proposed model has been found efficient and accurate in nearly all of the domains, except Recall, where the proposed model results are slightly lower than the existing model.

The proposed model has been found nearly 5-10 percent accurate than the existing model on the basis of precision results. The proposed model has been consistently found with 100% precision, which shows total absence of the false positive cases. It means the proposed model does not include any of the extra words while extracting the news data from the target source.

Table 4-5: Precision based comparison of proposed & existing models for ‘The Hindu

News ID	Existing Model (Precision)	Proposed model(Precision)
1	95	100

2	91	100
3	95	100
4	95	100

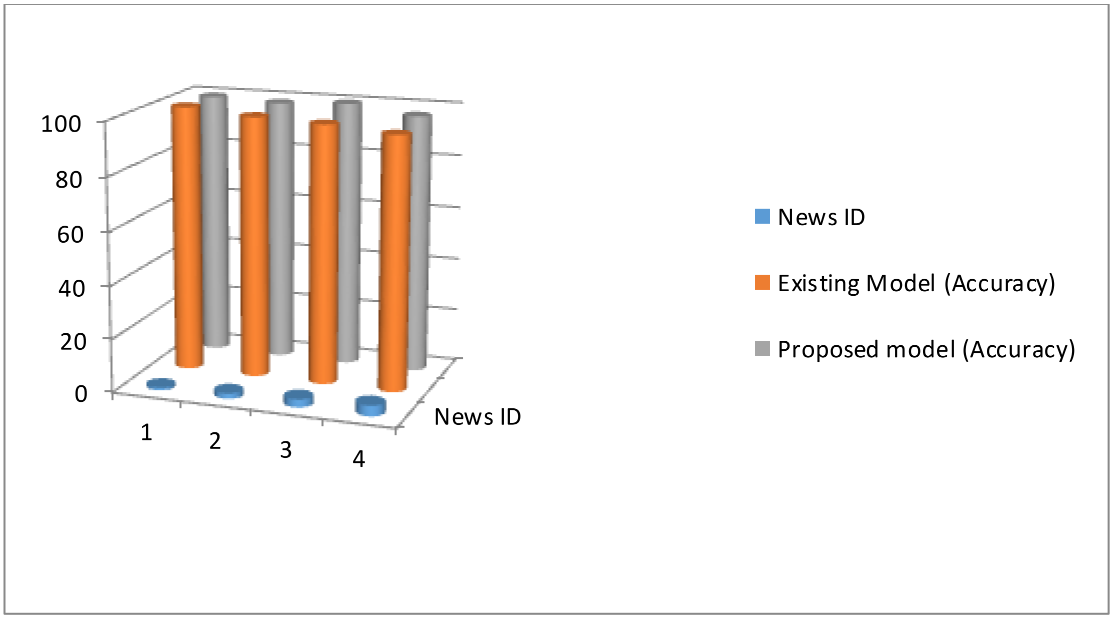


Figure 4-4: Precision based comparison of proposed & existing models for 'The Hindu'

The proposed model has been recorded at the slightly lower side, when compared on the basis of the recall value. This shows the existence of the false negative cases, which means some of the words are not getting extracted properly, which causes the recall value of 99.66%, which is nearly 0.34% lower than the existing model. But it provides the consistency of the work, where it varies the results from news to news.

Table 4-6: Recall based comparison of proposed & existing models for "The Hindu'

News ID	**Existing Model (Recall)**	**Proposed model (Recall)**
1	100	99.66
2	100	100

3	100	99.66
4	100	99.66

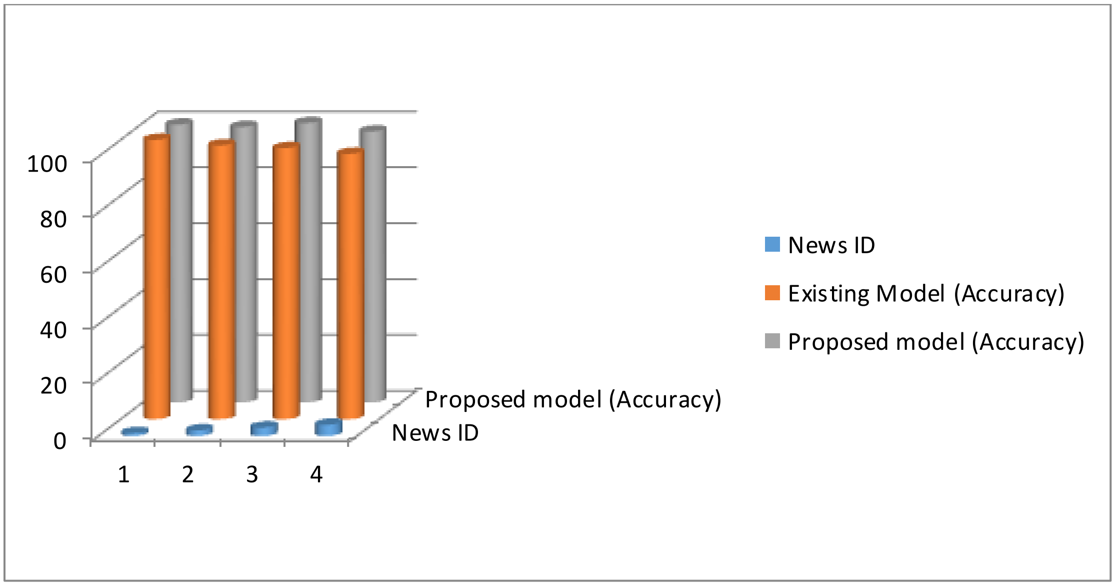


Figure 4-5: Recall based comparison of proposed & existing models for 'The Hindu'

The proposed model has been strongly found better than the existing model in the case of F1-measure, where the proposed model has been found nearly 3-5 percent better on all of the iterations in comparison to the existing model for the extraction and cleaning of the news data from "The Hindu" portal. As purposed method provides the better results and more than the existing model hence it improves the f1-measure values.

Table 4-7: F1-measure based comparison analysis of proposed & existing models for 'The Hindu'

News ID	**Existing Model (F1)**	**Proposed model (F1)**
1	97	99.83
2	95	100
3	96	99.83

4	96	99.83

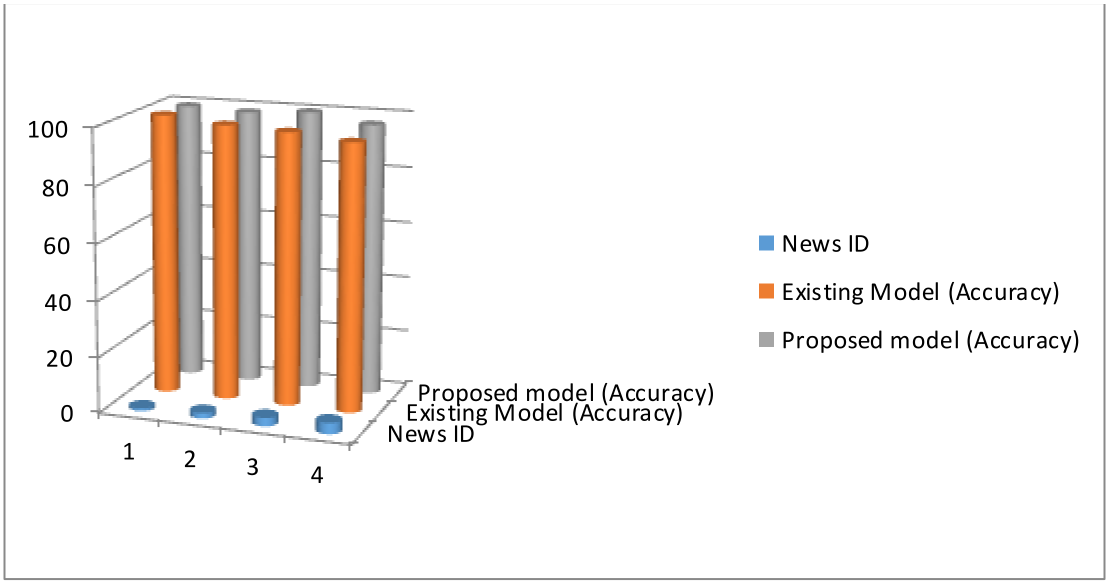


Figure 4-6:F1-measure based comparison of proposed & existing models for 'The Hindu'

The proposed model has posted the strong results in case of the overall accuracy, where the proposed model has been found nearly 4-9 percent better on all of the iterations in comparison to the existing model for the extraction and cleaning of the news data from "The Hindu" portal.existing model provides the accuracy 91- 95 % where as purposed model provides accuracy ranges from 99.66 to 100%.

Table 4-8: Accuracy based comparison of proposed & existing models

News ID	Existing Model (Accuracy)	Proposed model (Accuracy)
1	95	99.66
2	91	100
3	95	99.66
4	95	99.66

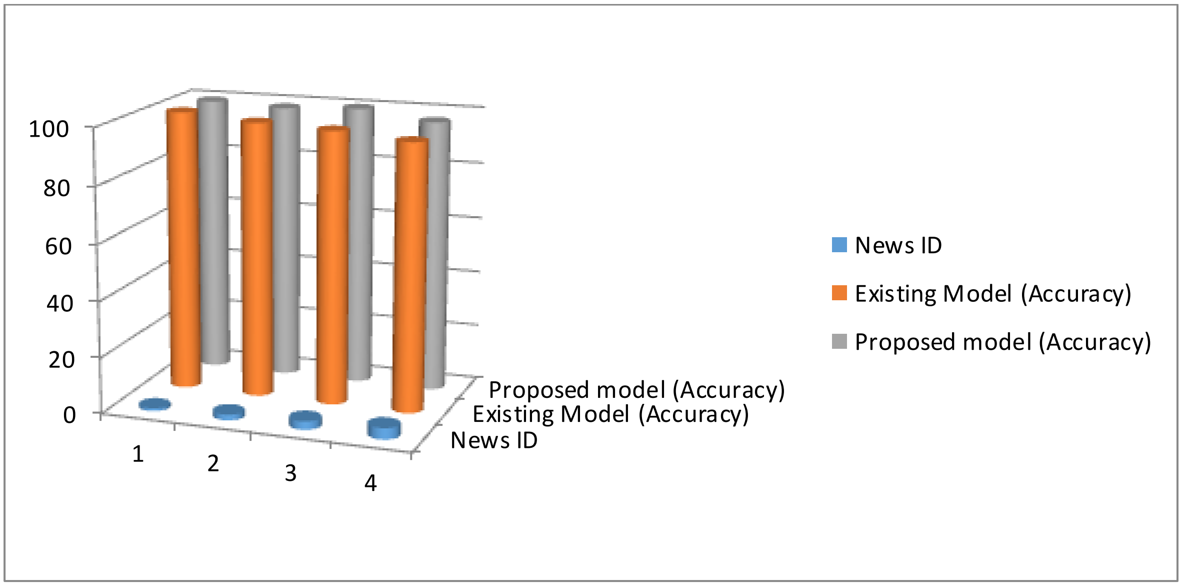


Figure 4-7: Accuracy based comparison of proposed & existing models for 'The Hindu'

Precision of the existing model when applied on the ‘indian Express’ is varied from 90 to 100%.Where as in purposed model precision 98.6 to 100%.Hence it shows the improved performance of the purposed model.

Table 4-9: Precision based comparison of proposed & existing models for 'Indian Express'

News ID	Existing Model (Precision)	Proposed model (precision)
1	100	100
2	98	98.6
3	98	100
4	96	99.18
5	90	99.88

6	95	100
7	100	99.52

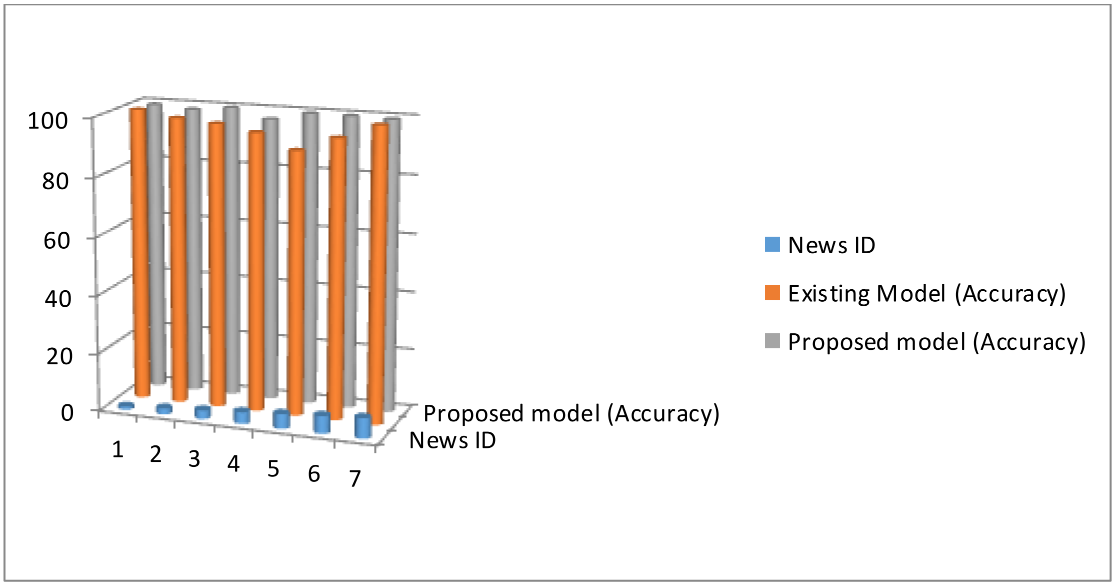


Figure 4-8: Precision based comparison of proposed & existing models for 'Indian Express'

Recall of the existing model for Indian express results are 98 to 100%.On the other hand purposed model provides recall 99.84 to 100%.Hence it gives more efficient results than the existing model.

Table 4-10: Recall based comparison of proposed & existing models for 'Indian Express'

News ID	Existing Model (Recall)	Proposed model (Recall)
1	100	100
2	100	100
3	98	99.46
4	98	99.59

5	100	100
6	100	100
7	100	99.78

Table shows recall based results where it gives the clear view of improvement to the existing model through purposed model. It depicts the lower performance of the existing model on the basis of recall values. Purposed model enhances the performance on the basis of recall values as it gave better results for the ‘Indian Express(I E)’ news text.

Graph clearly depicts the changes of the performance where it increases from the existing model. Both existing model and purposed models results for recall values shown on the graph. In this way bar graph provide the effective of the results to identify the performance issues.

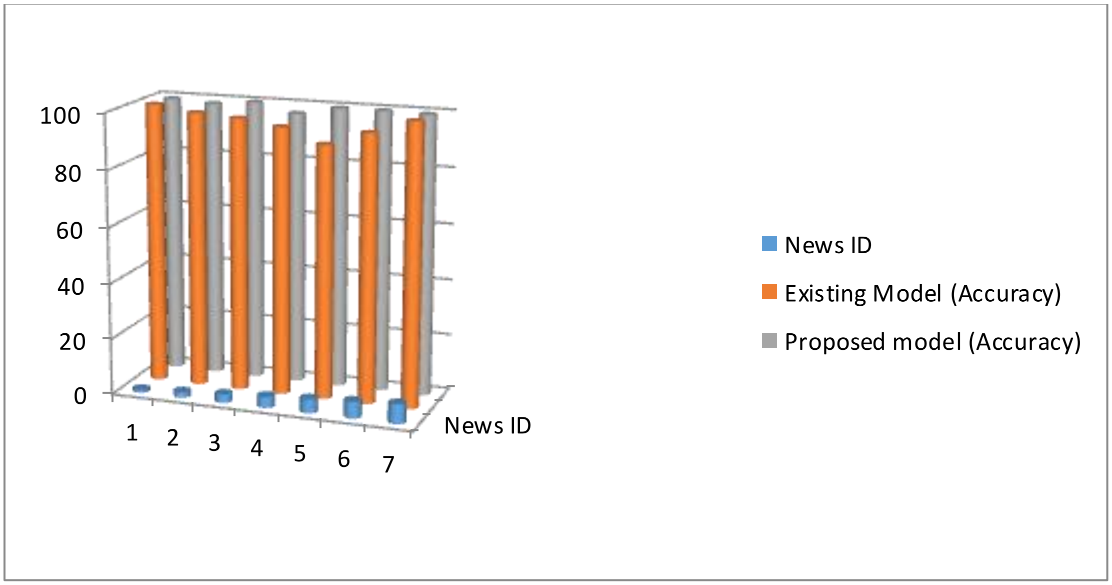


Figure 4-9: Recall based comparison of proposed & existing models for 'Indian Express'

Existing model provides the result 94 to 100% whereas purposed model provide the F1 results from 98 to 100%.Apart from that purposed mode also gives consistent results whereas existing model results varies from 94% to 97% and 98%.Hence purposed model provides more efficient results for F1.

Table 4-11: F1-Measure based comparison of proposed & existing models for 'I E’

News ID	Existing Model (F1)	Proposed model (F1)
1	100	99.73
2	98	99.2
3	98	100
4	97	99.38
5	94	99.89
6	97	99.76
7	100	99.60

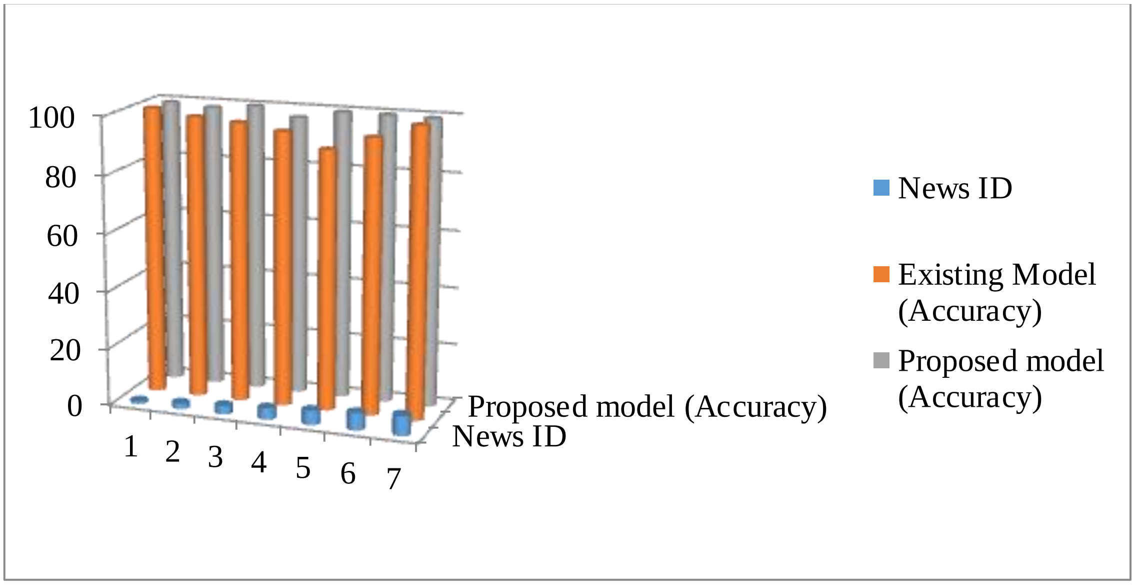


Figure 4-10: F1-Measure based comparison of proposed & existing models for 'IE'

Table 4-12: Accuracy Based Comparison of existing and purposed models for 'I E’

News ID	Existing Model (Accuracy)	Proposed model (Accuracy)
1	100	99.46
2	98	98.6
3	97	100
4	95	96.98
5	90	99.76
6	95	99.78
7	100	99.52

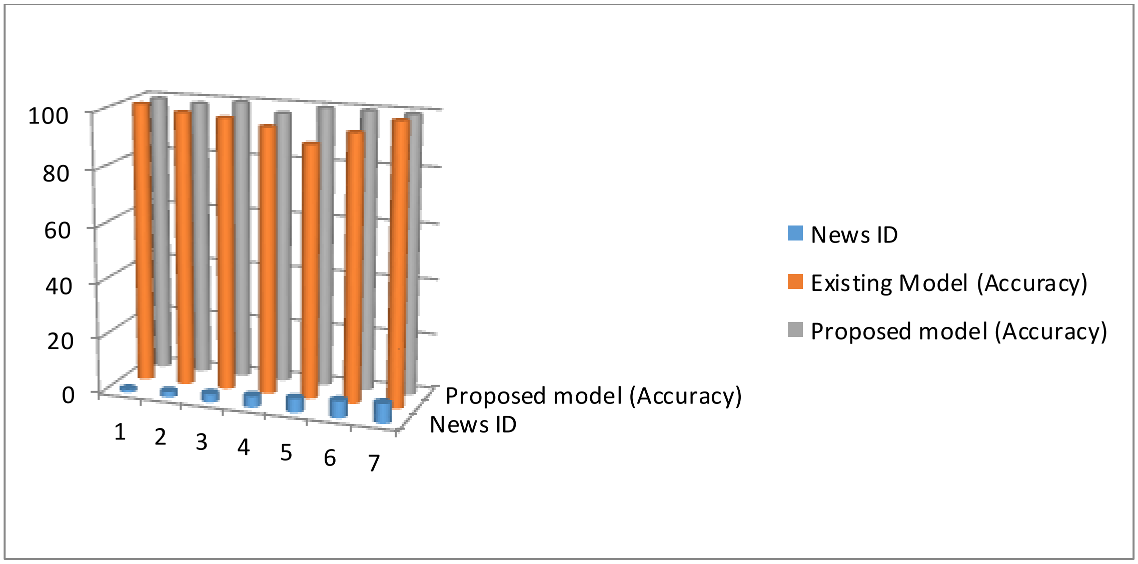


Figure 4-11: Accuracy Based Comparison of existing and purposed models for 'IE'

CHAPTER 5

CONCLUSIONS AND FUTURE SCOPE

Content extraction is the process to extract the important aspect of web pages and eliminate the extra unnecessary elements. To removal of noisy elements of the web pages several cleaning techniques are used. Hence user gets only required part rather than unrelated contents.

5.1 CONCLUSION

Web maintains the large amount of data which may be structured or unstructured.web pages contained the noisy elements so to provide efficient access to the required part ,it is important to avoid these noisy contents.

Web cleaning consists of various methods such as extracting meaningful contents and eliminating noise from the web pages. Purposed model is used to remove the noise from the internet newspaper and provide the particular news. Extraction of the specific news is based on the Dom tree. Purposed model consist the extraction of DOM tree So that useful content can be extracted. In this way target entities such as Main container can be extracted. At last it provides the main content without any noisy content. It actually scraps the title, brief introduction of the news and main part which is always inside the body. Div tags consist the paragraph tags which are mainly consisting the required part of the news or we can say the only news part. So the actual purpose is to clean the noise of the target news web page.

5.2 FUTURE SCOPE

As future work one can perform this model to the entire internet newspaper web site which may take more complex structure. Performance can also be improved. As web pages consists various extra elements such as images and advertisements. But some images may be important related to the contents; in this case a model can be developed to extract the images as well with required text.

REFERENCES

[1] Patel Niral, “A Survey on Web Mining with Noise Removal Technique,” *International Journal of Advavnce Research. Engineering ,Science Technology*, vol. 3, no. 5, pp. 443–447, 2016.

[2] D. Sahu and Y. Chouhan, “Comparative Study and Analysis on the Techniques of Web Mining,” *International Journal of Advance Research Computation and Communication Engineering*, vol. 5, no. 8, pp. 116–118, 2016.

[3] R. Kosala and H. Blockeel, “Web mining research:A Survey,” *ACM SIGKDD Explororation. Newsletter*, vol. 2, no. 1, pp. pp:1–15.

[4] S. Vijiyarani and M. E. Suganya, “Research issues in web mining,” *International. Journal of Computer aided Technology*, vol. 2, no. 3, pp. 55–64, 2015.

[5] R. C. Pereira and T. Vanitha, “Web Scraping of Social Networks,” *Int. J. Innov. Res. Commun. Eng.*, vol. 3, no. 7, pp. 237–240, 2015.

[6] S. Lanka, “A Comparative Study on Web Scraping,” *Proceeding 8th iternational Res. Conf.*, no. November, pp. 135–140, 2015.

[7] S. Gupta, G. Kaiser, D. Neistadt, and P. Grimm, “DOM-based content extraction of HTML documents,” *Proc. twelfth Int. Conf. World Wide Web WWW 03*, p. 207, 2003.

[8] D. Yang and J. Song, “Web content information extraction approach based on removing noise and content-features,” *Proc. - 2010 Int. Conf. Web Inf. Syst. Mining, WISM 2010*, vol. 1, pp. 246–249, 2010.

[9] H. Wang and Y. Zhang, “Web Data Extraction Based on Simple Tree Matching,” *WASE Int. Conf. Inf. Eng.*, pp. 15–18, 2010.

[10] S. López, J. Silva, and D. Insa, “Using the DOM tree for Content Extraction,” *Proc. 8th Int. Work. Autom. Specif. Verif. web Syst. 12)*, 2012.

[11] Y. F. Tseng and H. Y. Kao, “The mining and extraction of primary informative blocks and data objects from systematic Web pages,” *Proc. - 2006 IEEE/WIC/ACM Int. Conf.*

Web Intell. (WI 2006 Main Conf. Proceedings), WI'06, pp. 370–373, 2007.

[12] L. Yi, B. Liu, and X. Li, "Eliminating noisy information in Web pages for data mining," *Proc. ninth ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '03*, p. 296, 2003.

[13] S. Paek and J. R. Smith, "Detecting image purpose in World-Wide Web documents," *Proc. SPIE/IS&T Symp. Electron. Imaging*, 1998.

[14] A. Paniker, S. Panicker, P. Ravkhande, N. Tamboli, and P. R. Puntambekar, "Noise Reduction in Web Pages Using Featured DOM Tree .," *Int. J. Adv. Found. Res. Comput.*, vol. 3, no. 1, pp. 1–7, 2016.

[15] E. Silambarasan and S. S. Abbirammi, "Removal of Malicious Side Information in Web Documents," *Int. J. Contemp. Res. Comput. Sci. Technol.*, vol. 2, no. 3, pp. 551–554, 2016.

[16] D. Insa, J. Silva, and S. Tamarit, "The Journal of Logic and Algebraic Programming Using the words / leafs ratio in the DOM tree for content extraction <," *J. Log. Algebr. Program.*, 2013.

[17] M. R. Kaddu, "pages by using Hybrid Approach," *Int. Conf. Electr. ,Electronics Optim. Tech.*, 2016.

[18] J. Alarte, D. Insa, J. Silva, and S. Tamarit, "Web Template Extraction Based on Hyperlink Analysis," *.Proceeding XIV jornadas sobre Program. y Lenguajes(PROLE 15)Eptcs*, vol. 4204173, no. 102, pp. 16–26, 2015.

[19] F. Lei, M. Yao, and Y. Hao, "Improve the performance of the webpage content extraction using webpage segmentation algorithm," *IFCSTA 2009 Proc. - 2009 Int. Forum Comput. Sci. Appl.*, vol. 1, pp. 323–325, 2009.

[20] G. S. Yin, G. D. Guo, and J. J. Sun, "A template-based method for theme information extraction from web pages," *ICCASM 2010 - 2010 Int. Conf. Comput. Appl. Syst. Model. Proc.*, vol. 3, no. Jccasm, pp. 721–725, 2010.

[21] J. Kang, J. Yang, and J. Choi, "Repetition-based web page segmentation by detecting

tag patterns for small-screen devices," *IEEE Trans. Consum. Electron.*, vol. 56, no. 2, pp. 980–986, 2010.

[22] Y. L. Hedley, M. Younas, A. James, and M. Sanderson, "Sampling, information extraction and summarisation of Hidden Web databases," *Data Knowledge and Engineering*, vol. 59, no. 2, pp. 213–230, 2006.

[23] D. Pan, S. Qiu, and D. Yin, "Web page content extraction method based on link density and statistic," *2008 Int. Conf. Wirel. Commun. Netw. Mob. Comput. WiCOM 2008*, pp. 1–4, 2008.

[24] E. Uzun, H. Volkan, and T. Yerlikaya, "A hybrid approach for extracting informative content from web pages," *Information Process. Manag.*, vol. 49, no. 4, pp. 928–944, 2013.

[25] H. J. Carey and M. Manic, "HTML Web Content Extraction Using Paragraph Tags," *Int. Electronics. Electrical Engineering*, pp. 1099–1105, 2016.

[26] S. Sirsat and S. S. Science, "Pattern Matching for Extraction of Core Contents from News Web Pages," *Second International Coneference on Web Research*, pp. 13–18, 2016.

[27] K. Zhao, Y. Wang, X. Hu, and C. Wang, "Effective Blog Pages Extractor for Better UGC Accessing," *3rd International. Conference on Information Science and Control Engineering*, 2016.

[28] D. K. Mahto and L. Singh, "A Dive into Web Scraper World," *International Electronics. Electrical Engineering.*, pp. 689–693, 2016.

[29] R. Abarna and S. Pradeepa, "A Hybrid Approach for Extracting Web Information," *Indian Journal of. Science and Technology.*, vol. 8, no. August, pp. 1–6, 2015.

[30] S. B. K. G. A. Patil, "extracting brief note from Internet Newspapere," *International Electronics Electrical Engineering*, vol. 7, pp. 401–406, 2016.