"KEYWORD BASE NEWS FEED EXTRACTOR"

[1]KAMLESH S. SHARMA
P.G Department of Computer Science & Technology, D.C.P.E, H.V.P.M, Amravati, India
kamleshsatyanarayansharma@gmail.com
[2]KRUTIKA G. UPHADE
P.G Department of Computer Science & Technology, D.C.P.E, H.V.P.M, Amravati, India
krutikauphade14@gmail.com
[3]PROF. V. S. BELSARE
P.G Department of Computer Science & Technology, D.C.P.E, H.V.P.M, Amravati, India
vsbelsare@gmail.com

ABSTRACT: *User search for required information using search engine. In search engine various news portals are available on different news website. But user is interested only in the informative contents that he wants in various news portals. Also the non-content blocks such as sidebars, advertisements etc. can distract the user. Hence user needed to separate the informative content from non-informative content. Researchers will going to develop this system for remove various noise patterns in news portals. This system will help to provide different services according to the need of users. .*

***Keywords*:** Web mining, DOM extractor, search engine.

1. INTRODUCTION

The World Wide Web is the single largest data source in the world. This information which continues to expand in size and complexity .In which various news portals is available on the site. These news portals contain the noises that could affect the performance. Whenever user search the news it will return the many links related to the search. Now if the user has only two links related to user hence rest of all links is unwanted material for user. The World Wide Web has rich source of tremendous information which continues to expand in size and complexity. [1] Some of the data mining techniques applied in Web mining are association rule mining, clustering, classification, frequent item set. Some tasks of Web mining are finding of related information form selection of information and preprocessing, generalization and analysis.[1] Hence it will return only the user related search information in various news portals. The rest of materials which are not related to search are discarded.

2. LITERATURE WORK

There are many research paper based on content mining etc which are also related to the news feed system like.

[1] In this paper author proposed a hybrid approach to extract the main content from Web pages. A HTML Web page is converted to DOM tree and features are fetched and with the extracted features, rules are generated. For rules generation author used a Decision tree classification and Naïve Bays classification are machine learning methods. By using the rules, noisy part in the Web page is removing and informative content in the Web page is extracted. The performance of both decision tree classification and Naïve Bayes classification are measured with metrics like precision, recall, F-measure and accuracy.

[2]In this paper author proposed While HTML DOM analysis and visual layout analysis approaches have sometimes been used .it required higher accuracy in content extraction, the analyzing software needs to act a human user and under-stand content in natural language also eliminate noisy content. In this paper, Author describes a combination of HTML DOM analysis and Natural Language Processing (NLP) techniques for automated extractions of main article with associated images from web pages.

[3]In this paper, author focus on removing noise and utilization of all kinds of content-characteristics, experiments show that this approach can increase the universality and accuracy in fetching the body text of web pages.

3. WEB MINING

"Web mining refers to the overall process of unknown information or knowledge from the web data." Web mining is to apply data mining techniques to extract and uncover knowledge from web documents.

Web Mining Categories

Web mining may be divided into three categories:-

- Web usage mining
- Web structure mining
- Web content mining

Web Usage Mining

Web usage mining is the process of extracting useful information from server logs e.g. use Web usage mining is the process of finding out what users are looking for on the Internet.

Web Structure Mining

Web Structure Mining tries to focus on inter – document structure that is, to discover the link structure of hyperlinks.

.

Web Content Mining

Web Content Mining refers to the discovery of useful information from the web content. Here Content Refers to Text, etc. that numerous websites are holding.

4. SYSTEM OF DESIGN

- The propose system can be divided into 4 stages :-1)input keyword,2)comparing word and then link extract,3)remove the noisy information,4)provide link related to the keyword.
- Normally news portal contains the number of web pages and user is only interested in some news but in current system user need to go each and every page. To remove the time consumption and get accurate result oriented output we would like to develop the system in user which user input the keyword to search.
- In DOM extractor various news portals are available, through which user can select a specific news portal.
- The first stage in which user provide the input keyword.
- The second stage which compare the input keyword with each and every news heading.
- The third stage removes noisy information like advertisements, header, footer etc. by using DOM extractor (Data Object Module).
- The fourth stage when keyword is not match in DOM extractor they does not show result, if keyword is match they provide all related links .

Keyword based web news extractor algorithm

Step1:-login/registration of the user
Step2:-user selects a news paper.
Step3:-enter a keyword.
Step4:-user selects specific paper.
Step5:-than it goes to DOM extractor
Step6:-it match keyword with each headline in news paper which are present tags like <h1>, <h2>etc.
Step7:-remove unwanted news, advertisements etc.
Step8 (A):- if the keyword is identified then the results are displayed based on the proposed mechanism.
Step8 (B):- else
Result not found.
Step9:- show all related links to the keyword
Step10:-click on multiple links for complete information related to link.
Step11:-Stop

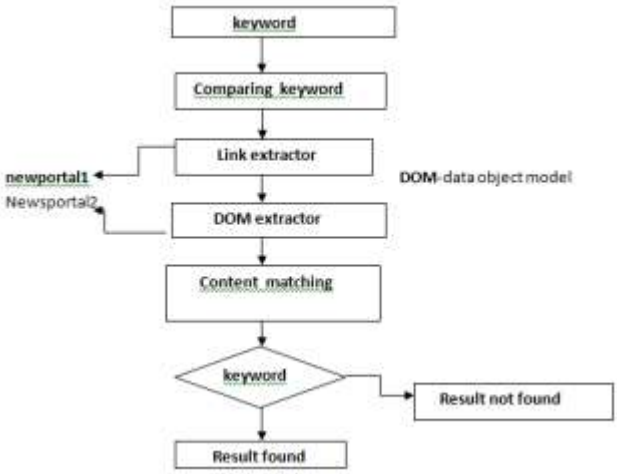

Figure 1: flow diagram for web news extractor

5. DOM TREE

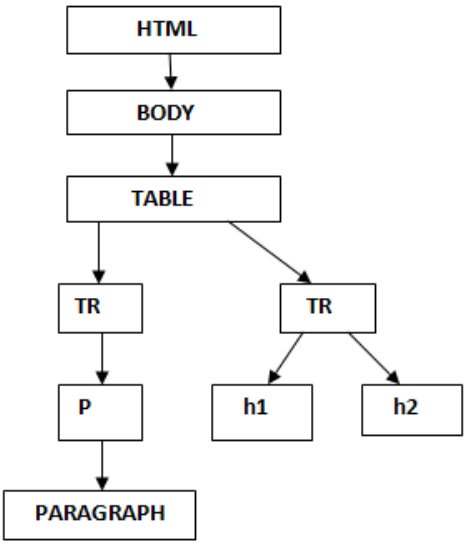

Figure 2: Dom extractor use in news feed system

.

Take the input web page for content extraction. After that pass the webpage through HTML parser that converts into HTML code. Now Create Document Object Model (DOM) tree for above HTML code.

The DOM is means that Document Object Model which is a standard for creating and manipulating in memory representation of HTML content. It defines logical structure of document and the way a document is accessed and manipulates.

The DOM tree is use for shorting the noisy information to valuable information. Apply various algorithms on DOM tree for extracting informative content. Finally we get desired output requested by user.

6. ADVANTAGES

- Filtration of noise from web pages.
- Finding appropriate output according to the user need.
- Complexity of web page is reduces.
- It is a user friendly approach.
- Time saving.

7. DISADVANTAGES

- Internet connection is must.
- When keyword is not match they don't show result.

8. CONCLUSIONS

Researchers will develop the system for extracting the target information from web pages. There are various news portals are available. It will separate the various news available with compare each and every news with user input.

Also web pages consist of noisy information like links advertisement, header, footer etc in news portals. Hence it will reduce or separate the original information form noisy information and provide appropriate output to user. News is efficiently extracted from various news portals.

9. REFERENCES

[1] "WEB CONTENT EXTRACTION USING HYBRID APPROACH". By K. Nethra, J. Anitha and G. Thilagavathi

[2] "Web Document Text and Images Extraction using DOM Analysis and Natural Language Processing" by Parag Mulendra Joshi, Sam Liu.

[3] "A New Approach for Web Information Extraction", byDr.S.Karthikeyan, R. Gunasundari.

[4] "Automating Content Extraction of HTML Documents", by Suhit Gupta, Peter Grimm.