An Efficient Web Content Extraction from Large Collection of Web Documents using Mining Methods

S. Mahesha
Research Scholar
CMJ University
Shillog

M. Giri, PhD.
Professor & Head
Department of CSE, SITAMS
Chittoor, Andhra Pradesh, India

M.S. Shashidhara, PhD.
Professor & Head
Department of MCA, OEC
Bangalore, India

ABSTRACT

Web mining is a one class of data mining. Web Mining is a variation of data mining that distills untapped source of abundantly available free textual information. The need and importance of web mining is growing along with the massive volumes of data generated in web day-to-day life. Web data Clustering is the organization of a collection of web documents into clusters based on similarity. A good clustering algorithm should have high intra-cluster similarity and low inter-cluster similarity. The process of grouping similar documents for versatile applications has put the eye of researchers in this area. In general, web data always arrives in a continuous, multiple, rapid and time varying flow. The Researchers in web mining proposed many methods to extract web contents, but they are fail to handle dynamic data. Web content extraction algorithms are important to extract useful contents from web sources. We propose a new method for web content extraction. It consist of four phases: Web document selection phase, web cube creation phase, web document preprocessing phase, and presentation phase. In the first phase list of web documents are selected for mining, second phase documents are used to create web cube, third phase documents are preprocessed, in the final phase results are presented to users. The experimental results of proposed system are compared with existing methods, Performance of proposed system is better than previous methods.

Keywords

Web Cube creation, Maintenance, Web document Cleaning, Web Mining

1. INTRODUCTION

The Web is undoubtedly the biggest repository of information in the history of humanity. According to recent estimates the contemporary World Wide Web contains up to 25 billion1indexed documents, available at over one trillion unique URLs and exposed by over 131 million unique domains. Electronic commerce uses to a large extent automated processes, resulting in the large availability of information in a digital form. Alternatively a high level of automation means that electronic commerce is especially dependent on information. Extraction of useful information is essential in the decision making process. Taxonomy of web mining is shown in figure 1 and which is broadly classified into three categories.

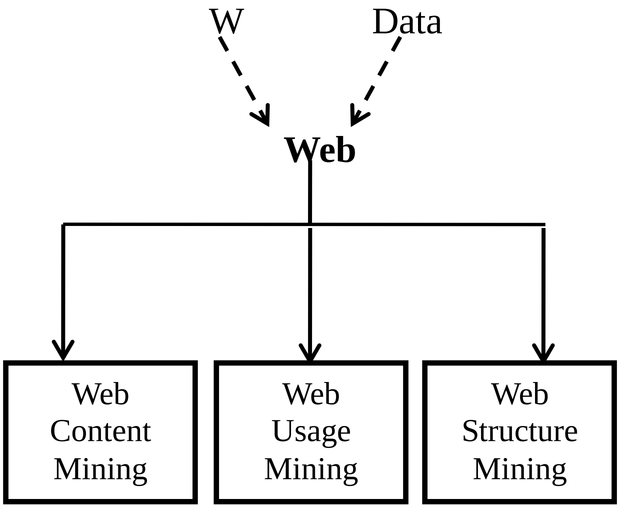


Figure 1: Taxonomy of web mining

Web Content Mining:

Web content mining use web data contents in the mining process. The general open question is how to mine content from web? In web content mining is the type of data mining techniques for relational databases, in the same manner we can expect to find similar types of knowledge extracted from unstructured data residing in web documents. The nature of web data is structured, semi-structured and unstructured, which forces a different approach towards web content mining. The web document contains a different type’s data such as text data, rich text data, image data, audios and video's, PPT, etc.

In this paper we primarily focus on mining useful information from the web content data. In particular, we consider the following issues of web content mining in the web information repositories context: In relational databases, the data are structured and are very well arranged in table using set of attributes and rows. In case of web repositories, web documents are unstructured. It is not that much easy to apply data mining system directly to search the entire WWW to discover required knowledge based on user query. In web content mining a list of web documents are selected from WWW to select useful information.

In our model, we mine based on key words, which are present in the documents. In this model we apply following steps:

First, a list of web documents related to one field is selected.

Second step, after selecting the documents, all documents are applied to cleaning method before mine effectively. Before mining, transform unstructured document to structured document, which is well understood.

All HTML tags in the documents are identified, after that only contents are extracted from tags. Likewise HTML Tags and its contents are separated, and then all HTML tags are removed from the documents, i.e. only key words are present in the document. After removing tags then weakly relevant or irrelevant documents are removed from selected list. Even the document is relevant and statements are irrelevant is also removed. Filter all irrelevant documents, statements, and words.

Finally create one XML document with document ID, word name, and word count. For each word in all preprocessed documents one row is added into XML file. Likewise, all unstructured and semi-structured documents are converted to structured document.

Third step, from XML document identify K-frequent keyword set and by using this keyword list, all association rules are extracted.

Fourth step, discovered rules are displayed to user.

Web Usage Mining:

Web usage mining is process of discovery of user access profile patterns from web server logs, which maintain history of each user when browsing the internet. In sever maintain logs containing information about the each user profile, list of pages accessed, time of accessing pages, user interested pages etc. This kind of information is used to maintain web site in an effective and efficient manner. Also finding the path from (Uniform Resource Locator) URL to last URL, associated list of web sites visited.

Web Structure Mining:

Information retrieval system in WWW use only the text on pages but ignoring valuable information contained in external links pages. The main aim of Web structure mining is to generate structural summary report about web sites and web pages. The main focus of web structure mining is therefore on external link information, which is very important aspect of web data. From the collection of selection of interconnected web documents, and discover more interesting and informative facts describing their connectivity of web subset. We are very much interested in generating structural information from the web rows stored in the web tables.

Measure the frequency count of the local links in the web rows in a web table. Local links mean connecting two documents, which reside on the same server. This informs about the web rows between connected documents in the web table that have more information about inter-related documents residing at the same server. This is also measure the completeness of the web sites in a sense that most of the closely related information is available at the same site.

Interior links means links which are within the same document. Also measure the frequency count of web rows in the web table containing interior links. This measure on a web document is able to identify cross-reference on other related web pages within the same document. The same measure is used to find the flow of the web documents. This information gives the relevant information is available within the same document.

Global links means links span over different web sites• Also measure the frequency count of web rows in the web table containing global links. This measure of the web documents is able to relate similar or related documents over different sites.

A huge amount of information is stored in WWW (World Wide Web). It contains an enormous and valuable content of textual or multimedia form; a significant part of Web information is semi-structured and contains data of business value. “Financial crisis of India” e-commerce is an example of a domain where number of Web documents exposing semi-structured (HTML documents) information exists, including manufacturers Web documents which containing the specification and classification of sales statistics, offered products, and catalogue prices and financial status. Suppliers, competitors and auction Web documents which containing valuable and continuously evolving information on product assortment, availability, prices, user reviews and delivery options. User opinion and professional product assessment Web documents which containing user ratings and different kinds of benchmark measurements as well as search engines and different types of social Web documents which providing information on product and brand popularity.

Acquiring all categories of data cited above is a critical task for most of market participants in many branches of the contemporary economy, as well as due to a high number of diverse and quickly evolving data sources on the Web, it is a complex research problem. As a result, the task of accessing semi-structured information available on the Web, called Web content extraction, has been an active field of both research and business Intelligence. The previous work in the area of content extraction from Web documents covered a few specific problems, including data extraction from static documents, learning extraction rules based on training data, user interaction or similarity of multiple documents, and acquisition of data from Deep Web sources that is Web information repositories hidden behind query forms.

Most of these problems have been already studied in depth, and well-performing methods exist. However, the WWW continued to evolve and brought in new content extraction challenges. Over time, many of Web documents have started to use composite Web documents using (i)frames and JavaScript to combine multiple HTML document, complex server interaction paradigms, advanced user interface elements based on a wide use of JavaScript and Flash components as well as on the rich event model and different types of stateful design things. The problem of extracting contents from Web documents using these technologies and interaction models further referred to as complex Web documents, which is central to this paper has not been addressed by any previous methods.

The remaining sections of the paper are structured as follows. We begin by describing the problem statement and objectives of the paper in section 2. In section 3, we present a new architecture of proposed system. In section 4, we discuss experimental setup of our proposed system. In section 5, discuss related work. Finally, section 6 gives conclusions and direction of future work.

2. PROBLEM DESCRIPTION

The previous Researchers proposed many methods for extraction of information from World Wide Web. The Research paper studies a set of problems that are faced during web data extraction. Researchers in web proposed many methods to extract patterns from web search engines. In web most of the information present is useless. The list of challenges faced is given below.

- Adopting a flexible approach to choosing the most valuable data during extraction (utility maximization)
- Combining multiple extraction languages using different features
- Combining content-based and context-based extraction rules
- Connecting data from multiple pages into records
- Dealing with AJAX-like requests
- Working with composite (multi-frame, using JavaScript and CSS) Web content.
- Dealing with data duplicates
- Dealing with different style formats.
- Dealing with non-deterministic navigation patterns.
- Dealing with technical issues of HTTP(S) (Cookies, redirects, encryption, compression, error codes, and certificates).
- Dealing with timer-triggered Web actions.
- Supporting distributed data extraction.
- Enabling multiple alternative rules for augmented robustness and better maintenance
- Enabling user interaction during data extraction (e.g. in order to handle CAPTCHAs)
- Extracting data from complex data presentation structures.
- Extracting hierarchies of unknown depth.
- Implementing advanced, state-aware query planning with multiple optimization steps for both navigation and extraction.
- Limiting number and frequency of requests.
- Taking advantage of typical constructs of Web documents: pagination, in-built ordering and dynamic filters.
- Using additional server query capabilities.
- Using both encoded and user content for data extraction.
- Using incompletely specified rules.
- Working with interactive content (Web actions based on interaction with user interface)

In this paper we propose a new method which solves the above mentioned challenges and the problems like web noisy data, junk mails, spam mails, advertisements, etc. This method focuses on the following objectives:

- Focusing on the role of web content extraction and identifying list problems when mining list of documents. Studying the solutions to these problems.
- Presenting the method which is used to identify required patterns in an effective manner.
- Examining a number of available techniques that can be applied to discover by solving these problems to achieve better performance.

3. ARCHITECTURE

The research area of this paper concerns Web content extraction that is the extraction of useful data from semi-structured Web sources. Web content means a collection of web documents are used for mining and extracting useful contents from those documents. The research problem studied in this paper concerns Web content extraction from various types of Web documents that is Web documents that are stateful, use advanced server interaction patterns such as asynchronous communication, or rely on client-side dynamic technologies, including JavaScript, and frames in their user interfaces. While this problem shares many web documents with extensively studied problems of performing data extraction from static documents and Web sources, it is significantly more challenging issues. Only a few of the challenges specific for data extraction from complex Web documents have been previously partially addressed, and most of them have not been even explicitly defined in the previous work. Therefore, Web content extraction from complex data-intensive Web documents should be treated as a new research problem.

Our research objective is to propose a content extraction model and algorithms capable of performing content extraction from previously handled basic data-intensive Web documents and Web sources as well as from Web documents that were not handled by existing approaches and takes lot of time to extract contents from web documents. The proposed model and algorithms should be characterized by less time consuming by separating and filtering contents from embedded tags at a time from a collection of web documents.

The proposed solution consists of a few elements, and in our opinion well addresses the stated objective. The paper that we will defend in this dissertation is that using query planning and execution algorithms based on an extensible, state aware data extraction model and on a rich representation of Web documents, enables a less time consuming content extraction from a significantly larger set of Web documents when compared with previous solutions.

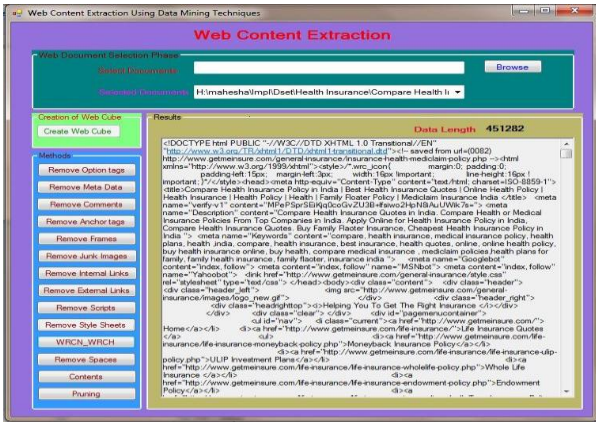
Architecture of proposed system is shown in figure 2. The main idea of proposed system is to extract patterns based on user interest using a collection of web documents by creating web cube. Architecture of proposed knowledge discovery from web databases includes following steps:

- Decide targeted data.
- Selection of input documents for mining.
- Apply Preprocessing techniques to clean web documents.
- Display contents to users.

In the proposed architecture, first a list of documents are selected and interesting patterns is fixed by the user by using interfaces. After collecting list of documents, all are applied to web data preprocessing step. In preprocessing step all list of selected documents are applied to cleaning, filtering and steaming process. Output of preprocessing is called content and it is displayed to user as knowledge.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

Proposed new web content extraction for mining web contents consist of four phases: Web document selection phase, web cube creation phase, Web document preprocessing phase, and Visualization phase. After selecting web documents related to one particular domain, all documents are merging and creating web cube, it is depicted in figure 3.



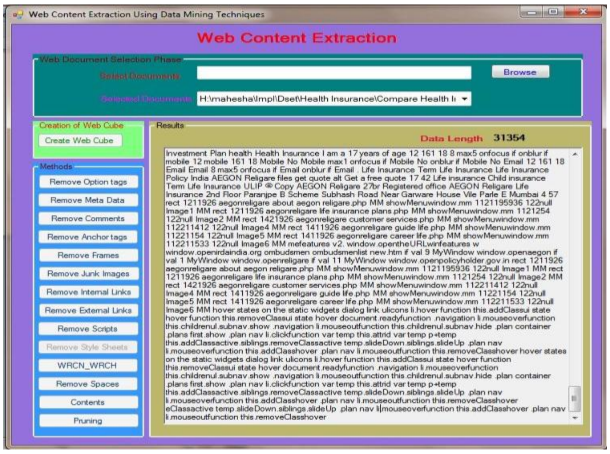
Selected Documents: H:\mahesha\Impl\Dset\Health Insurance\Compare Health I

Data Length 451282

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd"><!-- saved from url=(0082)
http://www.getmeinsure.com/general-insurance/insurance-health-mediclaim-policy.php --><html
xmlns="http://www.w3.org/1999/xhtml"><style>/*.wrc_icon{            margin:0;  padding:0;
           padding-left:15px;    margin-left:3px;       width:16px !important;          line-height:16px !
important; }*/</style><head><meta http-equiv="Content-Type" content="text/html; charset=ISO-8859-1">
<title>Compare Health Insurance Policy in India | Best Health Insurance Quotes | Online Health Policy |
Health Insurance | Health Policy | Health | Family Floater Policy | Mediclaim Insurance India </title>   <meta
name="verify-v1" content="MPePSprSEiKjq0coGvZU3B+lfsiwo2HpN8iAuUWik7s="> <meta
name="Description" content="Compare Health Insurance Quotes in India. Compare Health or Medical
Insurance Policies From Top Companies in India. Apply Online for Health Insurance Policy in India,
Compare Health Insurance Quotes. Buy Family Flaoter Insurance, Cheapest Health Insurance Policy in
India ">  <meta name="Keywords" content="compare, health insurance, medical insurance policy, health
plans, health ,india, compare, health insurance, best insurance, health quotes, online, online health policy,
buy health insurance online, buy health, compare medical insurance , mediclaim policies,health plans for
family, family health insurance, family flaoter, insurance india ">     <meta name="Googlebot"
content="index, follow"> <meta content="index, follow" name="MSNbot"> <meta content="index, follow"
name="Yahoobot">   <link href="http://www.getmeinsure.com/general-insurance/style.css"
rel="stylesheet" type="text/css"> </head><body><div class="content">    <div class="header">
<div class="header_left">                    <img src="http://www.getmeinsure.com/general-
insurance/images/logo_new.gif">                         </div>                  <div class="header_right">
             <div class="headrighttop"><i>Helping You To Get The Right Insurance </i></div>
            </div>       <div class="clear"> </div>       <div id="pagemenucontainer">
                         <ul id="nav">     <li class="current"><a href="http://www.getmeinsure.com/">
Home</a></li>       <li><a href="http://www.getmeinsure.com/life-insurance/">Life Insurance Quotes
</a>                    <ul>                                  <li><a href="http://www.getmeinsure.com/life-
insurance/life-insurance-moneyback-policy.php">Moneyback Insurance Policy</a></li>
                               <li><a href="http://www.getmeinsure.com/life-insurance/life-insurance-ulip-
policy.php">ULIP Investment Plans</a></li>                                          <li><a
href="http://www.getmeinsure.com/life-insurance/life-insurance-wholelife-policy.php">Whole Life
Insurance </a></li>                                          <li><a
href="http://www.getmeinsure.com/life-insurance/life-insurance-endowment-policy.php">Endowment
Policy</a></li>                                              <li><a
```

Figure 3: Creation of web cube from selected web documents

After applying all prepressing methods data size is gradually reduced. Extraction of contents, apply pruning method to prune extracted data, results are presented in result box, length of the cube is estimated and it is presented in data length filed, after this task size of the web cube is reduced, and it is depicted in figure 4.



Selected Documents: H:\mahesha\Impl\Dset\Health Insurance\Compare Health I

Data Length 31354

```
Investment Plan health Health Insurance I am a 17 years of age 12 161 18 8 max5 onfocus if onblur if
mobile 12 mobile 161 18 Mobile No Mobile max1 onfocus if Mobile No onblur if Mobile No Email 12 161 18
Email Email 8 max5 onfocus if Email onblur if Email . Life Insurance Term Life Insurance Life Insurance
Policy India AEGON Religare files get quote alt Get a free quote 17 42 Life insurance Child insurance
Term Life Insurance ULIP © Copy AEGON Religare 27br Registered office AEGON Religare Life
Insurance 2nd Floor Paranjpe B Scheme Subhash Road Near Garware House Vile Parle E Mumbai 4 57
rect 1211926 aegonreligare about aegon religare.php MM showMenuwindow.mm 1121195936 122null
Image1 MM rect 1211926 aegonreligare life insurance plans.php MM showMenuwindow.mm 1121254
122null Image2 MM rect 1421926 aegonreligare customer services.php MM showMenuwindow.mm
112211412 122null Image4 MM rect 1411926 aegonreligare guide life.php MM showMenuwindow.mm
11221154 122null Image5 MM rect 1411926 aegonreligare career life.php MM showMenuwindow.mm
112211533 122null Image6 MM mefeatures v2. window.opentheURLwinfeatures w
window.openirdaindia.org ombudsmen ombudsmenlist new.htm if val 9 MyWindow window.openaegon if
val 1 MyWindow window.openreligare if val 11 MyWindow window.openpolicyholder.gov.in rect 1211926
aegonreligare about aegon religare.php MM showMenuwindow.mm 1121195936 122null Image1 MM rect
1211926 aegonreligare life insurance plans.php MM showMenuwindow.mm 1121254 122null Image2 MM
rect 1421926 aegonreligare customer services.php MM showMenuwindow.mm 112211412 122null
Image4 MM rect 1411926 aegonreligare guide life.php MM showMenuwindow.mm 11221154 122null
Image5 MM rect 1411926 aegonreligare career life.php MM showMenuwindow.mm 112211533 122null
Image6 MM hover states on the static widgets dialog link ulicons li.hover function this.addClassui state
hover function this.removeClassui state hover document.readyfunction .navigation li.mouseoverfunction
this.childrenul.subnav.show .navigation li.mouseoutfunction this.childrenul.subnav.hide .plan container
.plans first.show .plan nav li.clickfunction var temp this.attrid var temp p+temp
this.addClassactive.siblings.removeClassactive temp.slideDown.siblings.slideUp .plan nav
li.mouseoverfunction this.addClasshover .plan nav li.mouseoutfunction this.removeClasshover hover states
on the static widgets dialog link ulicons li.hover function this.addClassui state hover function
this.removeClassui state hover document.readyfunction .navigation li.mouseoverfunction
this.childrenul.subnav.show .navigation li.mouseoutfunction this.childrenul.subnav.hide .plan container
.plans first.show .plan nav li.clickfunction var temp this.attrid var temp p+temp
this.addClassactive.siblings.removeClassactive temp.slideDown.siblings.slideUp .plan nav
li.mouseoverfunction this.addClasshover .plan nav li.mouseoutfunction this.removeClasshover
eClassactive temp.slideDown.siblings.slideUp .plan nav li.mouseoverfunction this.addClasshover .plan nav
li.mouseoutfunction this.removeClasshover
```

Figure 4: Pruning of extracted contents

After applying preprocessing techniques, the list of created web cube and its size is gradually decreasing by removing unnecessary contents, and it is depicted in Table1. The selected web document size is 454 KB and after removing unnecessary data extracted content size is only 32KB, which is only 10% of selected data. Therefore 90% of useless data is present in the document.

S.No	Method	Data Size
1	Web Cube Gen	454KB
2	Rem_Option	434KB
3	Rem_Meta	375KB
4	Rem_Anchor	198KB
5	Rem_Iframe	193KB
6	Rem_Image	184KB
7	Rem_Ilink	184KB
8	Rem_Elink	183KB
9	Rem_Script	119KB
10	Rem_Style	118KB
11	Rem_WRCN_WRCN	110KB
12	Rem_Spaces	86KB
13	Rem_Content	35KB
14	Rem_PContent	32KB

Table 1: Final summary of web content extraction

Selected web cube size is gradually decreased when applying our methods and it is shown in the figure 5. Finally in the existing method web content extractor display one page at a time, when a user wants to see other page again the user must select next page, but at a time get the data from multiple documents are not possible with these methods. In the proposed method if the user interested to we the data from multiple documents, then simply all the documents are selected and it applied to our method, then the proposed method mine all the documents and the contents from multiple documents are displayed to the user. So the proposed method takes less time to extract the contents when compared with the existing methods.

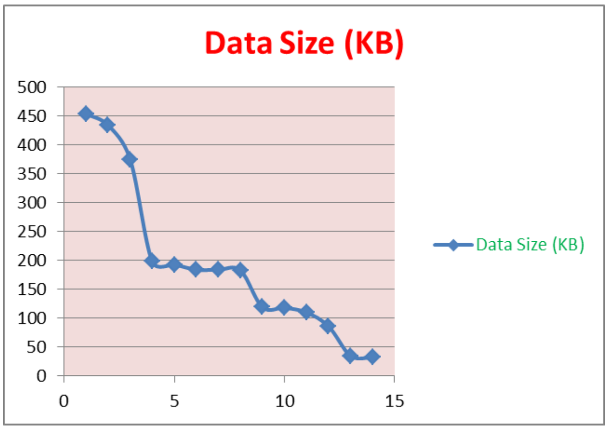


Figure 5: Experimental Results

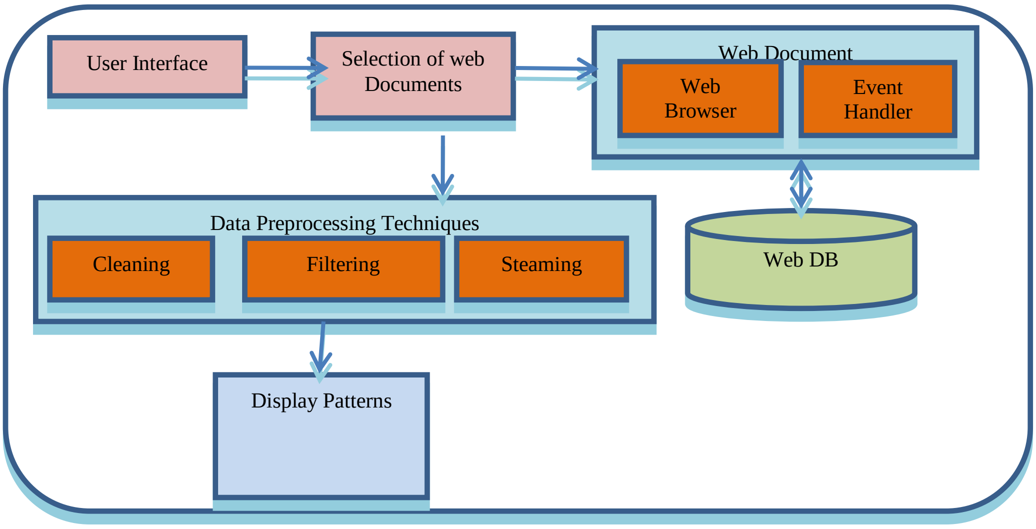


Figure 2: Architecture of proposed system

5. RELATED WORK

While recent years have seen a rapid growth of web multimedia content which including photos, audio, movies, Flash animations and 3D models, with Google web sites hosting over 240 million videos and with a YouTube large number of videos 2006-2008 growth rate of above 1900%, the vast majority of Web content consists of different kinds of textual documents with or without accompanying media, formatting, etc. They are provided in a number of different formats (HTML, PDF, XML, DOC, etc.) and vary from plain text to semi-structured documents containing data records. In case of HTML documents, that typically contain multiple media, an important area of Web pages is still covered by textual content [4].

An implementation of data preprocessing for web usage mining and the facts of algorithm for path completion are existing in Yan Li's paper [11]. After user session discovery, the missing pages in user access paths are append by using the referrer based method which is an effective solution to the problems introduce by proxy servers and local caching. The reference distance end to end of pages in complete path is modified by taking into account the average reference length of pages. As confirmed by practical web access log file, the path completion algorithm, proposed by Yan LI, efficiently appends the lost information and improves the reliability of contact data for further web usage mining calculations.

JIANG Chang-bin and Chen Li [12] bring about a Web log file data preprocessing algorithm based on collaborative filtering. It can make user session identification fast and flexibly even though statistical data are not enough and user history visiting records are absence. Huiping Peng [13] used FP-growth algorithm for processing the web log file records and obtained a set of frequent patterns. Then using the grouping of browse interestingness and site topology interestingness of association rules for web mining they revealed a new pattern to provide valuable data for the site construction.

In Web Usage Mining, web session data clustering plays vital role to classify visitors of website on the basis of user profile access history and similarity measure. Web session clustering is used in many ways to manage the web resources effectively such as personalization of web data, modification of schema. Dr. Sohail Asghar, Tasawar Hussain [14] proposed a method for web session clustering for preprocessing level of web usage mining. This method covers preprocessing steps to prepare the web log information and converts the unqualified web log data into numerical data.

Doru Tanasa[15], in his paper bring two significant contributions for a web usage mining. They proposed a complete methodology for preprocessing the Web logs and a divisive general methodology with three approaches for the discovery of sequential patterns with a low support. Ling Zheng[16], proposed improved data preprocessing to solve some existing problems in traditional data preprocessing technology for web log mining. Deep Web [5] (as opposed to Surface Web) is a huge part of the Web, consisting of data intensive Websites advertising their content only via query interfaces (rather than hyperlinks) [8, 10]. As its content remains mostly non-indexed by general purpose search engines, it is also referred to as Invisible Web [6, 7] or Hidden Web. HTML forms are the active components of Web pages used to acquire user input. In the case of Deep Web documents, user input is used to generate Web pages containing responses to a specific user queries.

6. CONCLUSION

The explosive day-to-day growth of information available on the web has necessity the web users to make use of some techniques to locate desired information from web resources. Web contains noisy data, redundant information and which mirrored web pages in and abundance. The effective way of identifying required patterns is a major issue the necessity to discover data from web sources and needs to be address. In this paper we propose an efficient method to address some of the problems during web content extraction. In the proposed method we extract required patterns by removing noise that is present in the web document.

We propose architecture framework, first a list of documents are selected and interesting patterns is fixed by the user by using interfaces. After collecting list of documents, all are applied to web data preprocessing step. In preprocessing step all list of selected documents are applied to cleaning, filtering and steaming process. Output of preprocessing is called content and it is displayed to user as knowledge. Then, we propose our own approach to information extraction from Web documents, consisting of models and methods as well as their implementations. Proposed method shows better performance when compared with existing methods. In future we plan to extend our work to construct DOM tree (Graphical representation) after extraction of useful patterns.

7. REFERENCES

[1] Magdalini Eirinaki and Michalis Vazirgiannis. Web mining for web personalization. ACM Transactions on Internet Technology, 3(1):1-27, 2003.

[2] Dimitrios Pierrakos, Georgios Paliouras, Christos Papatheodorou, and Constantine D. Spyropoulos. Web usage mining as a tool for personalization: A survey. User Modeling and User-Adapted Interaction, 6(2):311-372, 2003.

[3] Zhen Zhang. Large-scale deep web integration: Exploring and querying structured data on the deep web, 2006.

[4] Ryan Levering and Michal Cutler. The portrait of a common HTML web page. In 2006 ACM symposium on Document engineering, pages 198-204, 2006.

[5] Michael K. Bergman. The deep web: Surfacing hidden value. The Journal of Electronic Publishing, 7(1), 2001.

[6] King-Ip Lin and Hui Chen. Automatic information discovery from the 'invisible web'. In 2002 International Conference on Information Technology: Coding and Computing, page 332, 2002.

[7] Dirk Lewandowski and Philipp Mayr. Exploring the academic invisible web. Library Hi Tech, 24(4):529-539, 2006.

[8] Bin He and Kevin Chen-Chuan Chang. Statistical schema matching across web query interfaces. In 2003 ACM SIGMOD International Conference on Management of Data, 2003.

[9] Witold Abramowicz, Dominik Flejter, Tomasz Kaczmarek, Monika Starzecka, and Adam Walczak. Semantically enhanced deep web. In 3rd International AST Workshop, pages 675-680, 2008.

[10] Dominik Flejter and Tomasz Kaczmarek. Wybrane aspekty integracji informacji z g l,ebokiego Internetu, pages 97-110. Wydaw. AE, 2007.

[11] Yan LI, Boqin FENG and Qinjiao MAO, "Research on Path Completion Technique In Web Usage Mining", IEEE International Symposium On Computer Science and Computational Technology, pp. 554-559, 2008.

[12] JING Chang-bin and Chen Li, " Web Log Data Preprocessing Based On Collaborative Filtering ", IEEE 2nd International Workshop On Education Technology and Computer Science, pp.118-121, 2010.

[13] Huiping Peng, "Discovery of Interesting Association Rules Based On Web Usage Mining", IEEE Conference, pp.272-275, 2010.

[14] Tasawar Hussain, Dr. Sohail Asghar and Nayyer Masood, "Hierarchical Sessionization at Preprocessing Level of WUM Based on Swarm Intelligence ", 6th International Conference on Emerging Technologies (ICET) IEEE, pp. 21-26, 2010.

[15] Doru Tanasa and Brigitte Trousse,"Advanced Data Preprocessing for Inter-sites Web Usage Mining ", Published by the IEEE Computer Society, pp. 59-65, March/April 2004.

[16] Ling Zheng, Hui Gui and Feng Li, "Optimized Data Preprocessing Technology For Web Log Mining", IEEE International Conference On Computer Design and Applications(ICCDA), pp. VI-19-VI-21,2010.

[17] Nahm, U.Y., Bilenko, M. and Mooney R.J. "Two Approaches to Handling Noisy Variation in Text Mining". ICML-2002 Workshop on Text Learning, 2002

[18] Shian-Hua Lin and Jan-Ming Ho. Discovering Informative Content Blocks from Web Documents, KDD-02, 2002.

8. AUTHOR'S PROFILE

Mahesha S received MCA from Madurai Kamaraj University, Madurai. He received M. Phil in Computer Science from Vinayaka Missions University, Salem, Tamilnadu. He received PGDBC in Computer Science from Mysore University and obtained 6th Rank. He is pursing Ph. D from CMJ University, Meghalaya. He is presently working as Assistant Professor in the Department of Computer Science at Sree Siddaganga College of Arts, Science and Commerce for Women, Tumkur. He has around Ten years of teaching experience. He has published 2 papers in National conferences. His area of interest includes Data Mining.

Dr. M. Giri, received Ph. D in Computer Science and Engineering from University of Allahabad, Allahabad, Uttar Pradesh. He received M. Tech in Computer Science and Engineering from School of Information Technology (SIT), Jawaharlal Nehru Technological University (JNTU), Hyderabad. He received B. Tech in Computer Science and Engineering from Sree Vidyanikethan Engineering College, Jawaharlal Nehru Technological University (JNTU), Hyderabad. He is currently working as Professor & Head, Department of CSE, Sreenivasa Institute of Technology and Management Studies, Chittor Andhra Pradesh, India. He has around twelve years of teaching experience. He published 8 papers in International Journals, five papers in National Journals, six papers in International conferences, ten papers in National conferences. His area of interest includes Data Mining, Mobile Ad-hoc Networks, Artificial Intelligence and Grid Computing.