Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 6, June 2013, pg.145 – 150

RESEARCH ARTICLE

AN IMPLEMENTATION OF WEB PERSONALIZATION USING WEB MINING TECHNIQUES

V. Shanmuga Priya[1], S. Sakthivel[2]
[1]Department of computer science, Periyar University, TamilNadu, India
[2]Department of computer science, Periyar University, TamilNadu, India

[1] vspriyarjpm@gmail.com; [2] velsakthi810@gmail.com

Abstract— Web mining is a class of data mining. In order to relieve a "Data Rich but Information Poor" dilemma, Data Mining emerged. Web Mining is a variation of this field that distils untapped source of abundantly available free textual information. The importance of web mining is growing along with the massive volumes of data generated in web day-to-day life. In general, web data always arrives in a multiple, continuous, rapid and time varying flow. Most of the existing conventional algorithms fail while handling such dynamic data. Web data extraction algorithms are important in extracting useful documents from streaming on-line sources. We propose a new method for web data extraction. It has three phases. In the first phase list of web documents are selected, second phase documents are preprocessed, in the final phase results are presented to users. Experimental results are compared with existing methods. Performance of proposed system is better than existing methods.

Key Terms: - Data Mining; Web Mining; Similarity; Dissimilarity; Content Extraction; Filtering

I. INTRODUCTION

Web mining is the process of extracting interesting patterns from web information repositories. Web mining techniques are broadly classified into three categories: web usage mining, web structure mining, and web content mining. The different categories of web content mining is depicted in figure 1. In web content mining are classified into two categories, one is web page mining and other one is search results. In web page mining, different class of web data (Html, Xml, Text, and Multimedia) used to discover patterns directly from the web contents of web pages. In web search mining is intended to extract patterns from web search engines. Based on links used in web structure mining, it is classified into two types, one is internal and the other is external. Web structure mining intended to reveal the structure of web sites and how they are connected. Web usage mining go through server log files to extract patterns that reveals usage of website by the users.

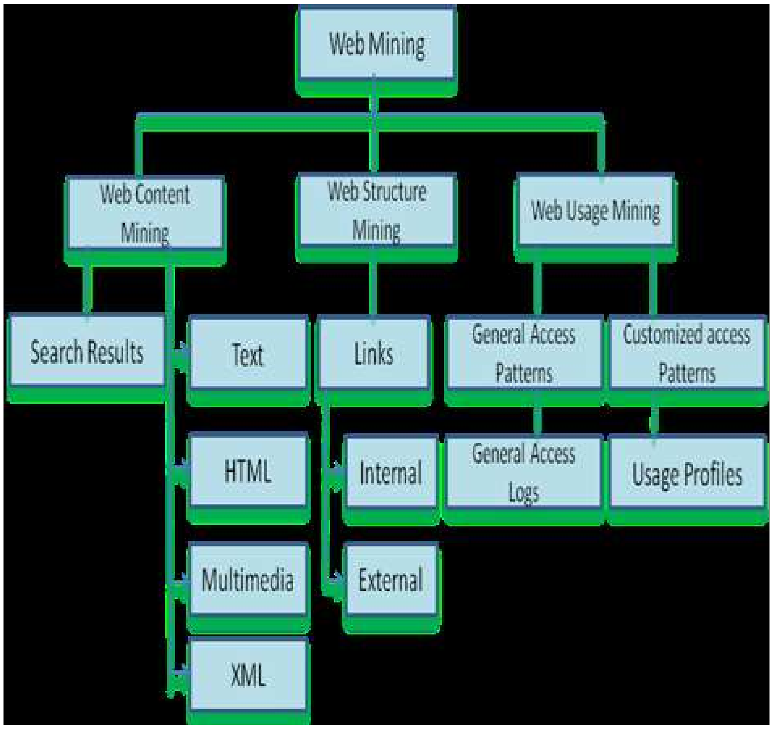


Figure1: Classification of Web Mining

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. Evolutionary clustering is an emerging research area which produces clusters that smoothly evolve over time. Evolutionary Clustering optimizes two potentially conflicting criteria: Snapshot Quality and History Cost simultaneously. In incremental clustering, the new clustering might not be related to the existing clustering whereas, evolutionary clustering is entirely based on the concept of maintaining the clustering over time.

A recent trend in clustering huge web data is the use of frequent item sets since they provide significant dimensionality reduction. In addition, frequent item sets address the problems like outlier removal, dimensionality reduction, etc. and satisfy the main features of evolutionary clustering, naturally and satisfy the two criteria of evolutionary clustering automatically.

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, we discuss related work. Finally, section 6 gives conclusions and direction of future work.

II. 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. In this paper we propose a new method which solves 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.

III. PROPOSED ARCHITECTURE

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.

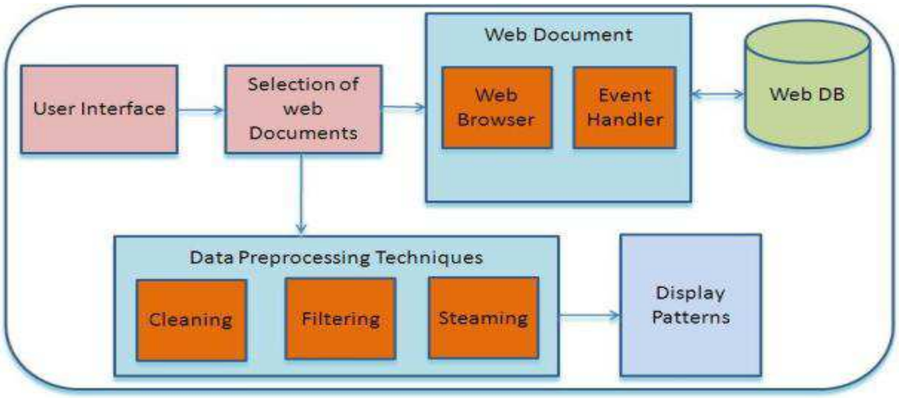
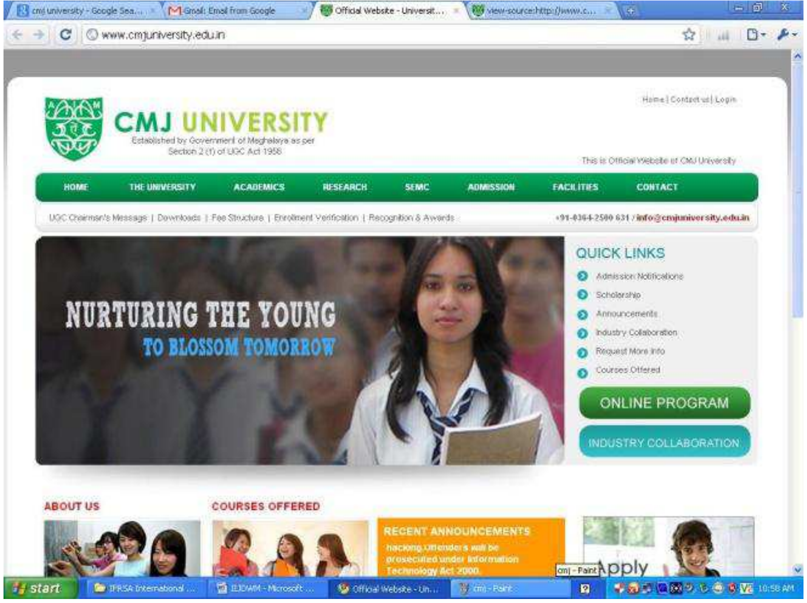


Figure 2: Architecture of proposed system

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.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

Web document presents data, for example user looking for CMJ university details then the site information is shown in figure 3 and site contents are shown in the figure 4. Fewer amounts of web data is presented using huge amount of tags.



Established by Government of Meghalaya as per Section 2 (f) of UGC Act 1956

This is Official Website of CMJ University

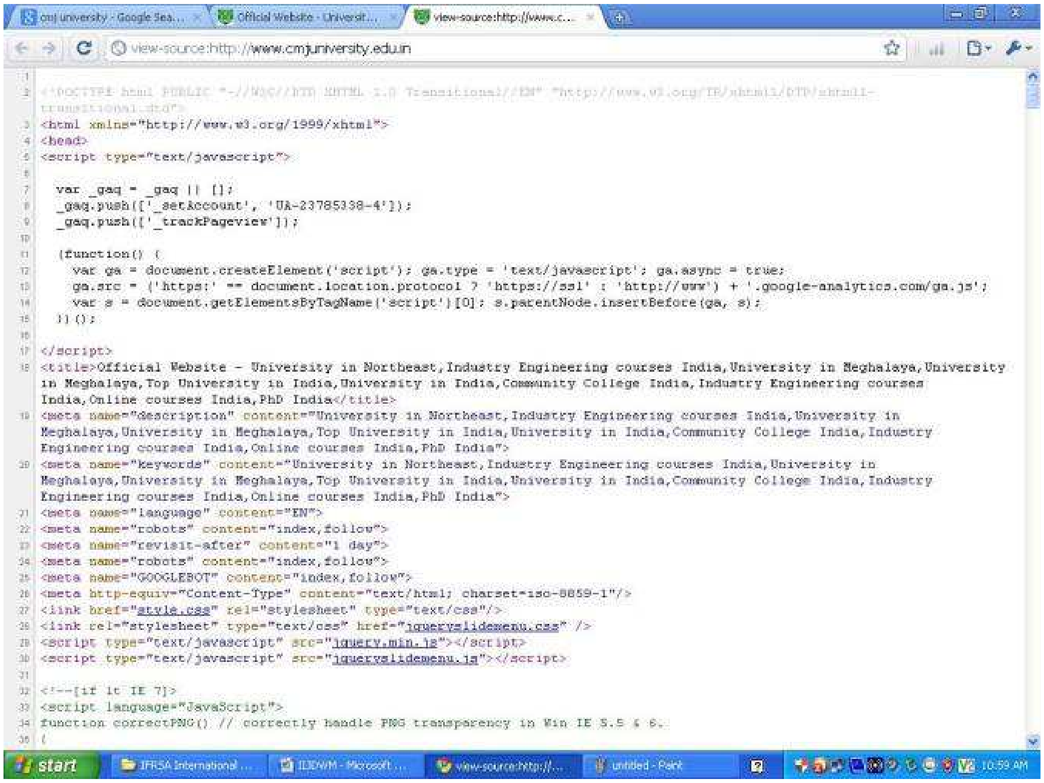
UGC Chairman's Message | Downloads | Fee Structure | Enrolment Verification | Recognition & Awards

+91-0364-2500 631 / info@cmjuniversity.edu.in

NURTURING THE YOUNG TO BLOSSOM TOMORROW

hacking.Offenders will be prosecuted under Information Technology Act 2000.

Figure 3: A sample Web Document


cmj university - Google Sea...
Official Website - Universit...
view-source:http://www.c...
view-source:http://www.cmjuniversity.edu.in
start
IFRSA International ...
IIJDWM - Microsoft ...
view-source:http://...
untitled - Paint
10:59 AM

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-
transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<script type="text/javascript">

  var _gaq = _gaq || [];
  _gaq.push(['_setAccount', 'UA-23785338-4']);
  _gaq.push(['_trackPageview']);

  (function() {
    var ga = document.createElement('script'); ga.type = 'text/javascript'; ga.async = true;
    ga.src = ('https:' == document.location.protocol ? 'https://ssl' : 'http://www') + '.google-analytics.com/ga.js';
    var s = document.getElementsByTagName('script')[0]; s.parentNode.insertBefore(ga, s);
  })();

</script>
<title>Official Website - University in Northeast,Industry Engineering courses India,University in Meghalaya,University
in Meghalaya,Top University in India,University in India,Community College India,Industry Engineering courses
India,Online courses India,PhD India</title>
<meta name="description" content="University in Northeast,Industry Engineering courses India,University in
Meghalaya,University in Meghalaya,Top University in India,University in India,Community College India,Industry
Engineering courses India,Online courses India,PhD India">
<meta name="keywords" content="University in Northeast,Industry Engineering courses India,University in
Meghalaya,University in Meghalaya,Top University in India,University in India,Community College India,Industry
Engineering courses India,Online courses India,PhD India">
<meta name="language" content="EN">
<meta name="robots" content="index,follow">
<meta name="revisit-after" content="1 day">
<meta name="robots" content="index,follow">
<meta name="GOOGLEBOT" content="index,follow">
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1"/>
<link href="style.css" rel="stylesheet" type="text/css"/>
<link rel="stylesheet" type="text/css" href="jqueryslidemenu.css" />
<script type="text/javascript" src="jquery.min.js"></script>
<script type="text/javascript" src="jqueryslidemenu.js"></script>

<!--[if lt IE 7]>
<script language="JavaScript">
function correctPNG() // correctly handle PNG transparency in Win IE 5.5 & 6.
{
```

Figure 4: Web document contents

Experimental setup of proposed system is shown in figure 5. Before web content extraction, first a list of web documents is selected. Then all selected documents are applied to data preprocessing technique. In web most of the data present is a noisy or dirty in the nature. If the contents are extracted from this type of documents gives incorrect patterns, so all the documents are applied to preprocessing techniques like cleaning, steaming and transforming.

During web content extraction process following methods are applied:

- **Remove Comments:** from the selected documents all comments are removed first.
- **Remove Meta tags:** Meta tags gives description about web documents, user interesting patterns are not in the description, so it removed.
- **Remove Scripts:** client side or server side scripting languages are used to present the document in a look and feel manner, so it can also remove from the documents.
- **Remove junk images or ads:** popup advertisements deviate user into web spoofing sites, so the image tags present in the document is removed.
- **Remove External or Internal links:** Link tags in the document are removed.
- **Cleaning:** less important words in the document like is, was, articles, etc from the documents.
- **Steaming:** synonyms are replaced with single word.
- **Transforming:** create web data cube.

Let suppose, before applying preprocessing web document size is 10KB, after applying preprocessing size is reduced to 1KB. Therefore 90% of data present in the document is noisy data which has been removed from the documents. According to user request extracted patterns are displayed to user as knowledge

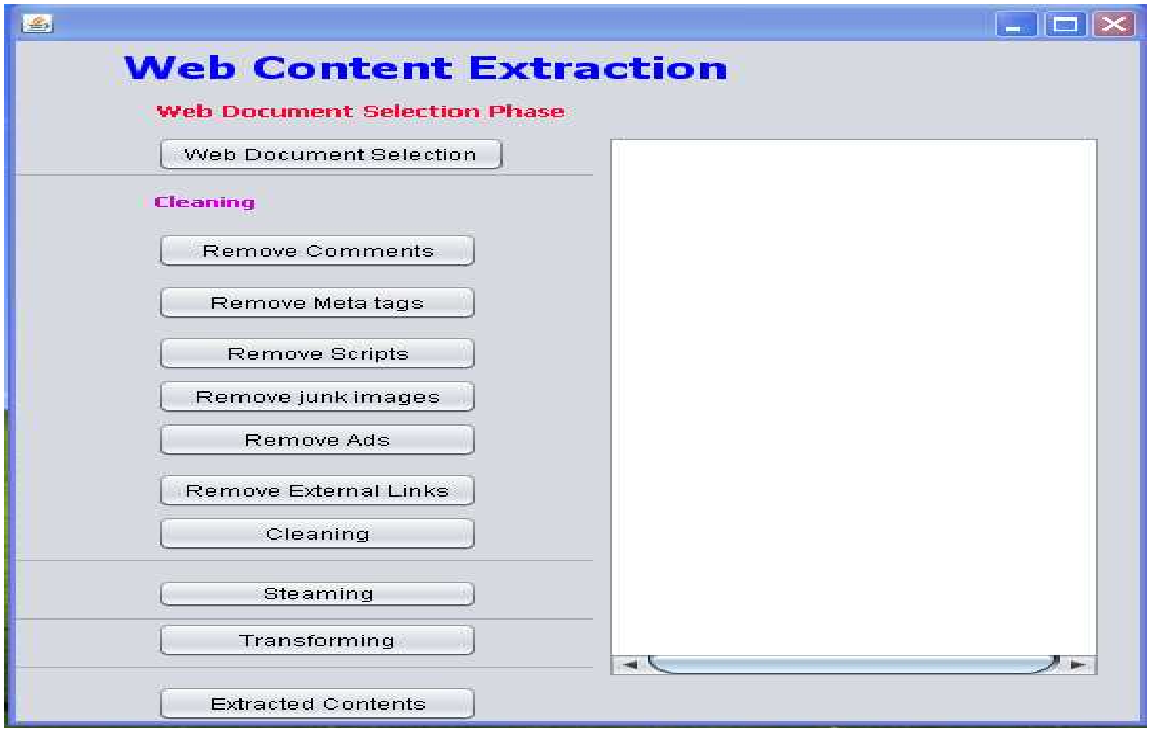


Figure 5: Experimental setup

V. Related Work

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 brings 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.

VI. Conclusion and Scope of Future Work

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. 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.

REFERENCES

[1] Dr. M. Giri and Dr. Akash Kumar, "An Efficient Web Content Mining using Relevance Analysis Approach", International Journal of Multidisciplinary Research in Advanced Engineering, pages. 201-210, 2012.

[2] Dr. M. Giri and Dr. Akash Kumar, "An Efficient Web Content Mining using Divide and Conquer Approach", International journal of Computational Intelligence Research, pages. 201-210, 2012.

[3] Dr. M. Giri and Dr. Akash Kumar, "An Efficient Web Content Mining using Multi Threading Approach", International Journal of Systems, Algorithms and Applications, pages. 1-4, 2012.

[4] Bettina Berendt, Andreas Hotho, Dunja Mladenic, Maarten van Someren, Myra Spiliopoulou, "A Roadmap for Web Mining: From Web to Semantic Web", Springer, 2005.

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

[6] Bar-Yossef. Z and Rajagopalan. S. Template Detection via Data Mining and its Applications, WWW, 2002.

[7] Cooley.R., Mobasher.B. and Srivastava.J, Data preparation for mining World Wide Web browsing patterns. Journal of Knowledge and Information Systems, (1) 1, 1999.

[8] Zhen Zhang; "Light-weight Domain–based Form Assistant: Querying web databases on the fly "; 31st VLDB Conference; Trondheim Norway; 2005.

[9] O. Zamir and O. Etzioni; "Web document clustering: a feasibility demonstration"; In Proceedings of SIGIR; 1998.

[10] Bin He, Kevin chen-chuan chang; "Statistical schema matching across web query interfaces"; In SIGMOD Conferences; 2003.

[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 Coference, 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 Intersites 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 OnComputer Design and Applications(ICCDA), pp. VI-19-VI-21,2010.