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. 4, April 2013, pg.79 – 83

REVIEW ARTICLE

A Review on Various Web mining Techniques with Purposed Algorithm of K-means Web Ranking

Mohinder Singh[1], Navjot Kaur[2]

[1]Student of masters of technology, Department of Computer Science and Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India

[2]Assistant Professor, Department of Computer Science and Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India

[1] *Bhullardimple@ymail.com;* [2] *Navjot_anttal@yahoo.co.in*

Abstract— With huge amount of information available online, the World Wide Web is fertile area of data mining research. The web mining research is crossroad of several forms of several research communities such as data base IR within AI, especially the subarea of Machine language and Natural Language processing. However there is lot of confusion with comparing research efforts from different point of views .In this Paper ,We Survey the research area of web mining ,point out confusion regarding the confusion of uses of term web mining and suggested three other web mining categories then we situate some of three categories. We also explore connection between web mining types and related agent paradigm. For the survey we focus on representation issue on the process and learning algorithm which is based on page ranking method. We conclude the paper with some research issues.

Key Terms: - web mining techniques; Page ranked Algorithm; Database; agent

I. INTRODUCTION

A. Web Mining [4]

Web mining is the application of data mining techniques to discover patterns from the web. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining. Here web content mining is to be used. Web content mining is related but different from data mining and text mining. It is related to data mining because many data mining techniques can be applied in Web content mining. It is related to text mining because much of the web contents are texts. However, it is also quite different from data mining because Web data are mainly semi-structured and/or unstructured, while data mining deals primarily with structured data. Web content mining is also different from text mining because of the semi-structure nature of the Web, while text mining focuses on unstructured texts. Web content mining thus requires creative applications of data mining and/or text mining techniques and also its own unique approaches. In the past few years, there was a rapid expansion of activities in the Web content mining area. This is not surprising because of the phenomenal growth of the Web contents and significant economic benefit of such mining. However, due to the heterogeneity and the lack of structure of Web data, automated discovery of targeted or unexpected knowledge information still present many challenging research problems. In this tutorial, we will examine the following important Web content mining problems and discuss existing techniques for solving these problems.

- **Data/information extraction**: Our focus will be on extraction of structured data from Web pages, such as products and search results. Extracting such data allows one to provide services. Two main types of techniques, machine learning and automatic extraction are covered.

- **Web information integration and schema matching**: Although the Web contains a huge amount of data, each web site (or even page) represents similar information differently. How to identify or match semantically similar data is a very important problem with many practical applications. Some existing techniques and problems are examined.
- **Opinion extraction from online sources**: There are many online opinion sources, e.g., customer reviews of products, forums, blogs and chat rooms. Mining opinions (especially consumer opinions) is of great importance for marketing intelligence and product benchmarking. We will introduce a few tasks and techniques to mine such sources.
- **Knowledge synthesis**: Concept hierarchies or ontology are useful in many applications. However, generating them manually is very time consuming. A few existing methods that explores the information redundancy of the Web will be presented. The main application is to synthesize and organize the pieces of information on the Web to give the user a coherent picture of the topic domain..
- **Segmenting Web pages and detecting noise**: In many Web applications, one only wants the main content of the Web page without advertisements, navigation links, copyright notices. Automatically segmenting Web page to extract the main content of the pages is interesting problem. A number of interesting techniques have been proposed in the past few years.

All these tasks present major research challenges and their solutions also have immediate real-life applications. The tutorial will start with a short motivation of the Web content mining. We then discuss the difference between web content mining and text mining, and between Web content mining and data mining. This is followed by presenting the above problems and current state-of-the-art techniques. Various examples will also be given to help participants to better understand how this technology can be deployed and to help businesses. All parts of the tutorial will have a mix of research and industry flavor, addressing seminal research concepts and looking at the technology from an industry angle.

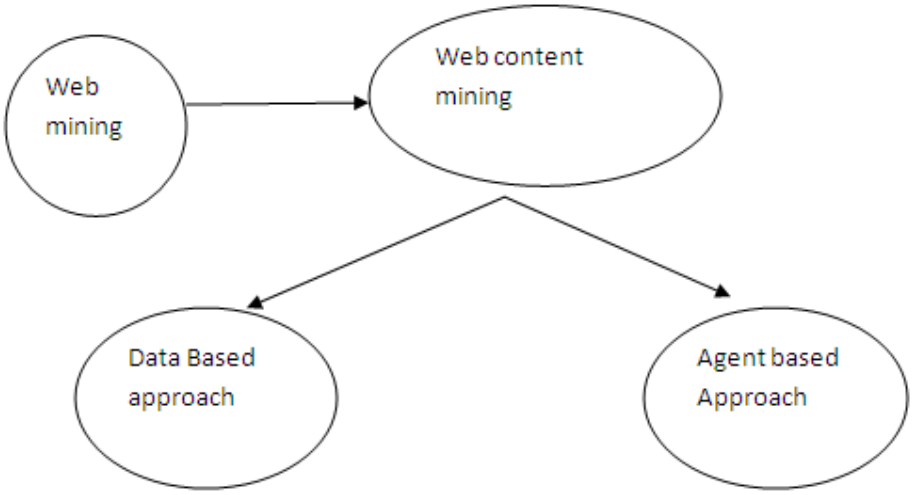

Figure 1 Types of web content mining

The database approaches to Web mining have generally focused on techniques for integrating and organizing the heterogeneous and semi-structured data on the Web into more structured and high-level collections of resources, such as in relational databases, and using standard database querying mechanisms and data mining techniques to access and analyze this information.

1. **Multilevel-Databases**

Several researchers have proposed a multilevel database approach to organizing Web-based information. The main idea behind these proposals is that the lowest level of the database contains primitive semi-structured information stored in various Web repositories, such as hypertext documents. At the higher level(s) meta data or generalizations are extracted from lower levels and organized in structured collections such as relational or object-oriented databases. For example, Han, et. al. use a multi-layered database where each layer is obtained via generalization and transformation operations performed on the lower layers. Kholsa, et. Al.propose the creation and maintenance of meta-databases at each information providing domain and the use of a global

schema for the meta-database. King & Novak propose the incremental integration of a portion of the schema from each information source, rather than relying on a global heterogeneous database schema. ARANEUS system extracts relevant information from hypertext documents and integrates these into higher-level derived Web Hypertexts which are generalizations of the notion of database views.

2. **WebQuery-Systems**

There have been many Web-based query systems and languages developed recently that attempt to utilize standard database query languages such as SQL, structural information about Web documents, and even natural language processing for accommodating the types of queries that are used in World Wide Web searches. We mention a few examples of these Web-based query systems here. W3QL combines structure queries, based on the organization of hypertext documents, and content queries, based on information retrieval techniques. WebLog Logic-based query language for restructuring extracted information from Web information sources. Lorel and UnQL query heterogeneous and semi-structured information on the Web using a labeled graph data model. TSIMMIS extracts data from heterogeneous and semi-structured information sources and correlates them to generate an integrated database representation of the extracted information.

3. **Agent-Based Approach**

The agent-based approach to Web mining involves the development of sophisticated AI systems that can act autonomously or semi-autonomously on behalf of a particular user, to discover and organize Web-based information. Generally, the agent-based Web mining systems can be placed into the following three categories:

a) **Intelligent-Search-Agents**
Several intelligent Web agents have been developed that search for relevant information using characteristics of a particular domain (and possibly a user profile) to organize and interpret the discovered information. For example, agents such as Harvest , FAQ-Finder , Information Manifold , OCCAM , and ParaSite rely either on pre-specified and domain specific information about particular types of documents, or on hard coded models of the information sources to retrieve and interpret documents. Other agents, such as ShopBot and ILA (Internet Learning Agent) , attempt to interact with and learn the structure of unfamiliar information sources. ShopBot retrieves product information from a variety of vendor sites using only general information about the product domain. ILA, on the other hand, learns models of various information sources and translates these into its own internal concept hierarchy.

b) **Information-Filtering/Categorization**:
A number of Web agents use various information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize them . For example, HyPursuit uses semantic information embedded in link structures as well as document content to create cluster hierarchies of hypertext documents, and structure an information space. BO (Bookmark Organizer) combines hierarchical clustering techniques and user interaction to organize a collection of Web documents based on conceptual information.

4. **Personalized-Web-Agents**

Another category of Web agents includes those that obtain or learn user preferences and discover Web information sources that correspond to these preferences, and possibly those of other individuals with similar interests (using collaborative filtering). A few recent examples of such agents include the WebWatcher, PAINT, Syskill & Webert , and others . For example, Syskill & Webert is a system that utilizes a user profile and learns to rate Web pages of interest using a Bayesian classifier

In this work the useful pattern will be extracted from web and will be shows as DOM (Graphical Representation) The Document Object Model (DOM) defines a standard way for accessing and manipulating HTML documents. The DOM presents an HTML document as a TREE-STRUCTURE. The DOM is a W3C (World Wide Web Consortium) standard. The W3C Document Object Model (DOM) is a platform and language-neutral interface that allows programs and scripts to dynamically access and update the content, structure, and style of a document.

II. PROPOSED ALGORITHM

A. PROBLEM DESCRIPTION

The previous Researchers proposed many methods for extraction of information from World Wide Web Paper studies a set of problems that are faced during extraction. Researchers in web proposed many methods to from web search engines. In web most of the information present is useless. In this paper we propose a new 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 identifying list problems when mining list of Studying the solutions to these problems.
- Presenting the method which is used to identify 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.

B. PROPOSED ARCHITECTURE

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.

III. 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 [10]. 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 lodistance 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 information and improves the reliability of contact data for further web usage mining calculations. JIANG Chang-bin and Chen Li [11] 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 used FP-growth algorithm for processing the web log file records and obtained a set of frequent patter Then using the grouping of browse interestingness and site pology 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 [13] 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.

IV. 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 S Shashidhara2, Dr. M. Giri*," An Efficient Web Content Extraction Using Mining Techniques", International Journal of Computer Science and Management Research, Vol 1 Issue 4 November 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] 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.
[7] Zhen Zhang; "Light-weight Domain–based Form Assistant: Querying web databases on the fly "; 31st VLDB Conference; Trondheim Norway; 2005.
[8] O. Zamir and O. Etzioni; "Web document clustering: a feasibility demonstration"; In Proceedings of SIGIR; 1998.
[9] Bin He, Kevin chen-chuan chang; "Statistical schema matching across web query interfaces"; In SIGMOD Conferences; 2003.
[10] 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.
[11] 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.
[12] Huiping Peng, "Discovery of Interesting Association Rules Based On Web Usage Mining", IEEE Coference, pp.272-275, 2010.
[13] 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.
[14] 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.

