ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

Analysis & Approaches of Web Mining

Rohit Sarla[1], Dr. R.C.Dixit[2], Dr, Pradeep Sharma[3] ,Prof. Rajesh Shah[4]

Research Scholar, Govt. Holkar Science College, Indore, M.P., India

Professor, Dept. of Physics, Govt. Holkar Science College, Indore, M.P., India

HOD, Dept. of CS, Govt. Holkar Science College, Indore, M.P., India

HOD, Dept. of CS & Elex., Christian Eminent College, Indore, M.P., India

ABSTRACT: The extracting the useful information from the web is very important task.web mining is the application of the data mining techniques.web mining is defined as the extract valuable information from the web.web mining is classified into three types.web content mining,web structure mining and web usage mining. To bring together and to present some of the latest research results in the field, to encourage more research activities in the field, with the huge amount of data/information already on the Web and more to come. The next big thing is naturally how to make best use of the Web to mine useful data/information and to integrate heterogeneous data/information automatically.

KEYWORDS: web mining, web content mining, web structure mining, web usage mining.

I.INTRODUCTION

The advancement in the technology covered faster communications. The previous decade experienced a dramatic development in computer technology, such that with the press of a finger the information about a particular topic appeared in monitors within seconds. As time passed by the complexity of web increased due to enormously large amount of data. So extraction of data according to users need became a tedious task. As a result mining became an essential technique to extract valuable information from internet and this technique was named as web mining. Web mining is further classified into three types which are Web content mining, Web Structure mining and Web Usage mining. Using the objects like text, pictures, multimedia etc. content mining is done in the web. In Web structure mining, mining is done based on the structure like hyperlinks. In the case of web usage mining, mining is done on web logs which contain the navigational pattern of users and the study of this navigational pattern will trace out the interest of the users [1].

II.OVERVIEW OF WEB MINING

Web mining means to discover the information from World Wide Web and it also find out its usage patterns .Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services.

Web mining should be decomposed into these subtasks:

1. The task of retrieving intended Web documents (Resource finding).
2. Automatically selecting and pre processing specific information from retrieved Web resources (Information selection and pre processing).
3. Automatically discovers general patterns at individual Web sites as well as across multiple sites (Generalization).
4. Validation and/or interpretation of the mined patterns (Analysis).

III.CATEGORIES OF WEB MINING

Web mining is categorized into three areas of interest based on part of Web to mine:

1. Describes discovery of useful information from contents, data and documents (Web content mining). Two different points of view: Information Retrieval view and Data Base view.
2. Model of link structures, topology of hyperlinks Categorizing of web pages (Web structure mining).
3. Mines secondary data derived from user interactions (Web usage mining).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

IV.WEB CONTENT MINING

Web content mining is the process of extracting useful information from the content of Web documents. Logical structure, semantic content and layout are contained insemi-structured Web page text. Topic discovery, extracting association patterns, clustering of Web documents and classification of Web pages are some of research issues in text mining. These activities use techniques from other disciplines – IR(information retrievel), IE (information extraction), NLP (natural language processing) and others. Automatic extraction of semantic relations and structures from Web is a growing application of Web content mining. In this area, several algorithms are used. Hierarchical clustering algorithms on terms in order to create concept hierarchies, formal concept analysis and association rule mining to learn generalized conceptual relations and automatic extraction of structured data records from semi-structured HTML pages. In contrast to unstructured texts, structured data is also easier to extract. This problem has been studied by researchers in Artificial Intelligence and database and data mining [3] [4].

V.WEB CONTENT MINING TECHNIQUES

It identifies the useful information from the web contents/data/documents, however, such a data in its broader form has to be further narrowed down to useful information. Web content data consist of structured data such as data in the tables, unstructured data such as free texts, and semi-structured data such as HTML documents. Here, the several approaches in web content mining are represented [5].
Web content mining becomes complicated when it has to mine unstructured, structured, semi structured and multimedia data.

1. Web content data is much of unstructured text data. The research around applying data mining techniques to unstructured text is termed Knowledge Discovery in Texts (KDT), or text data mining, or text mining. Hence one could consider text mining as an instance of web content mining to provide effectively exploitable results, preprocessing steps for any structured data is done by means of information extraction, text categorization, or applying NLP techniques. Content mining has been accomplished on unstructured data such as text. Mining of unstructured data provides unknown information. Text mining is extraction of previously unknown information by extracting information from different text sources. Content mining requires application of data mining and text mining techniques. Basic content mining is a type of text mining . Some of the useful techniques used in text mining are as Information Extraction, Information Visualization, Topic Tracking, Summarization, Categorization, and Clustering (Unstructured Data Mining Techniques).

2. The techniques which have been used for mining structured data are referred as Structured Data Mining (Structured Data Mining Techniques).
3. The techniques used for semi structured data mining are Object Exchange Model (OEM), Top down Extraction, and Web Data Extraction language (Semi-Structured Data Mining Techniques).
4. Some of the Multimedia Data Mining Techniques are SKICAT, Multimedia Miner, Color Histogram Matching and Shot Boundary Detection (Multimedia Data Mining Techniques) [6].

VI.WEB MINING TASKS

Web mining consists of the different essential tasks , which are described in a fig. below.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

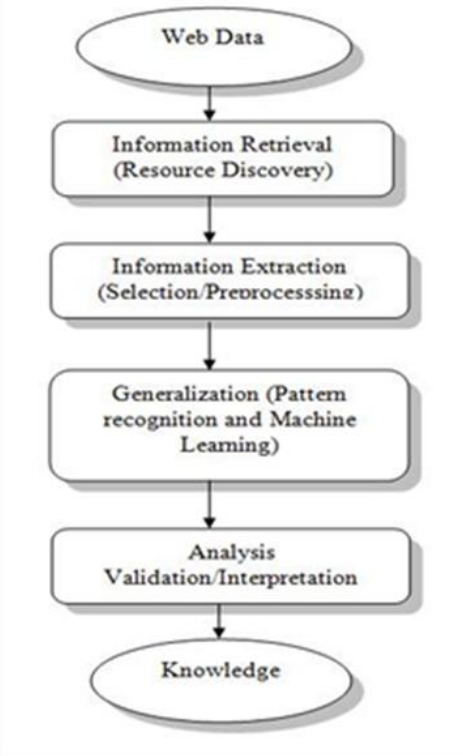

Fig. Web Task Mining

1. It is the task of retrieving the intended information from the Web. It locates the unfamiliar documents and services on the Web (Information Retrieval).
2. It is the task of automatically selecting and pre-processing specific information from retrieved Web resources (Pre-processing).
3. It is the task to automatically discover general patterns of individual Web sites as well as across multiple web sites (Pattern Recognition & Machine Learning).
4. It is the task of analyzing, validating and interpreting the mined patterns (Analysis) [7].

VII. PROBLEM STATEMENT

Users could encounter following problems when interacting with the Web.

a) Most people use some search service when they want to find specific information on the Web. A user usually inputs a simple keyword query and a result is a list of ranked pages. This ranking is based on their similarity to the query. Today's search tools have some problems are Low precision and low recall, mainly because of wrong or incomplete keyword query. This leads to irrelevance of many search results (Finding relevant information).

b) This problem is data-triggered process that presumes that we have a collection of Web data and we want to extract potentially useful knowledge from these data (Creating new knowledge).

c) People differ in the contents and presentations they prefer while interacting with the Web (Personalisation of information).

d) This is a group of sub-problems such as mass customizing information to intended consumers, problems related to effective Web site design and management, problems related to marketing and others (Learning about consumers or individual users) [8].

VIII. RELATED WORKS

Various scholars and researches have proposed related work in Web content mining, which are discussed below -

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

Aidan Finn discusses in his research work Fact or fiction, Content classification for digital libraries, methods for content extraction from single-article sources, where content is supposed to be in a single body. The algorithm tokenizes a page into either words or tags. The page is sectioned into three contiguous regions, placing boundaries to partition the document such that most tags are placed into outside regions and word tokens into the center region. This approach works well for single-body documents, but destroys the structure of the HTML and doesn't produce good results for multi-body documents, i.e., where content is segmented into multiple smaller pieces like we find on Web Blogs [9].

McKeon in the NLP (Natural Language Processing) group at Columbia University detects the largest body of text on a webpage (by counting the number of words) and classifies that as content. This method works well with simple pages. However, this algorithm produces noisy or inaccurate results handling multi-body documents, especially with random advertisement and image placement.

Rahman in first International workshop on Web Document Analysis propose another technique that uses structural analysis, contextual analysis, and summarization. The structure of an HTML document is first analyzed and then properly decomposed into smaller subsections. The content of the individual sections is then extracted and summarized. However, this proposal has yet to be implemented. Furthermore, while the paper lays out prerequisites for content extraction, it doesn't actually propose methods. Thus it again proves ineffective in actual content extraction. A variety of approaches have been suggested for formatting web pages to fit on the small screens of cellular phones and PDAs however, they basically end up only reorganizing the content of the webpage to fit on a constrained device and require a user to scroll and hunt for content. The main aim is however to device a method for the generic Web documents accessible on any device [9] [10].

IX.CONCLUSION

Architecture or module design for efficient web mining & make a coding for software if need for programming.Web mining uses various data mining techniques, but it is not an application of traditional data mining due to heterogeneity and unstructured nature of the data available on the World Wide Web. The main uses of web content mining are to gather, categorize, organize and provide the best possible information available on the WWW to the user requesting the information. The mining tools are imperative to scanning the many HTML documents, images, and text provided on Web pages. The resulting information is provided to the search engines, in order of relevance giving more productive results of each search detailed study and analysis of each web mining tools have been discussed. For future scope of web content mining includes predicting user needs in order to improve the usability, scalability and user retention.

REFERENCES

[1] Kosla, R. and Blockeel, H. 2000. Web Mining Research: A Survey. SIG KDD Explorations. Vol. 2, 1-15.
[2] G. Srivastava, K. Sharma, V. Kumar," Web Mining: Today and Tomorrow", in the Proceedings of 2011 3rd International Conference on Electronics Computer Technology (ICECT), pp.399-403, April 2011.
[3] Fan, W., Wallace, L., Rich, S. and Zhang, Z. 2005. Tapping into the Power of Text Mining. Communications of the ACM - Privacy and Security in highly dynamic systems. Vol. 49, Issue-9.
[4] Gupta,V. and Lehal,G.S. 2009.A Survey of Text Mining Techniques And Applications.Journal Of Emerging Technologies In Web Intelligence.Vol. 1,pp.60-76.
[5] Pol, K., Patil, N., Patankar, S. and Das, C. 2008. A Survey on Web Content Mining and extraction of Structured and Semi structured Data. IEEE First International Conference on Emerging.
[6] Dunham, M. H. 2003. Data Mining Introductory and Advanced Topics. Pearson Education.
[7] Han, J., Kamber, M.: Data Mining: Concepts and Techniques, Second edition, p. 628-648. Morgan Kaufmann Publishers, 2006.
[8] A. F. R. Rahman, H. Alam and R. Hartono. Understanding the Flow of Content in Summarizing HTML Documents. In Int. Workshop on Document Layout Interpretation and its Applications, DLIA01, Sep., 2001.
[9] A. F. R. Rahman, H. Alam and R. Hartono. Content Extraction from HTML Documents. In 1st Int. Workshop on Web Document Analysis (WDA2001), 2001.