**Web Mining And The Challenges**

Dr. A.Carmel Prabha, Technical Test Lead, Infosys Limited, Chennai.

ABSTRACT

This paper concentrates upon the current tendencies in Big Data analytics in the AWS clouds. How analytics on AWS Cloud altering the panorama of Big Data Analytics and how convenient it is for small organizations to bounce into Big Data Analytics space. Some mild is additionally thrown into the future scope of this concept.

**Keywords:**Big Data on Cloud, Big Data Analytics, Big Data Analytics on AWS Cloud, AWS Big Data tools, AWS QuickSite.

1. INTRODUCTION

“Necessity is the Mother of invention- Plato”

Data Mining refers to extraction of information from giant set of Data. It can additionally be referred to as Knowledge mining. As the Information Technology industry evolves rapidly with Data it is essential to turn such data into information, more precisely Knowledge. The WWW can be considered as the biggest database. Web mining is one of the very difficult and fast-evolving subjects in records mining. Because of the dynamic and unstructured nature of the internet data, the net information lookup has encountered a lot of troubles and challenges in the area of Data mining principles.

**Web Mining**

Web Mining is the term that involves the data mining principles to extract knowledge or pattern from World Wide Web document and services. Web mining is pretty exceptional from Data Mining as the former offerwith unstructured data(no pre-defined data model, mostly DOM structure where in every html tag corresponds to a node in DOM tree) and the latter with structured facts (Pre-defined and machine-readable, a locatable, on occasion relational ‘data model’ generally of real-world objects). Hence the Web mining can be used to enhance the power of search engines like Google in bringing out the appropriate web pages according to the search keywords, identifying authoritative web pages etc. The basic web mining task can be decomposed into the following subtask.

* Resource Discovery: Identifying the futureinfo from web.
* DataRemoval: Choosing and Pre-processing the infoon or after the above step automatically.
* Generalization: Involves automatically discovering individual patterns individually and also across multiple websites.
* Analysis: Analysing the generalized or mined pattern.

Taxonomy

Web Mining

Web Content Mining

Web Structure Mining

Web Usage Mining

1. Web Content Mining

Web content material mining (WCM) offers with mining, extraction integration of beneficial data, records and Knowledge from Web page contents.Today there are billions of net pages HTML documents, images and different multimedia documents handy on the Internetwhich are still continuously increasing. Taking into consideration of this factor it has become necessary to extract interesting contents from web. One of the best examples is Google’s web content mining with respect to our query. Let’s analyse the same in our case. The result of our search must match our requirements, technically speaking the web space must be personalized. As a trivial example a user from India who is looking for a computer or laptops should be redirected to the websites like Flipkart or Snapdeal from India similarly for the user from US should be directed to US websites like Amazon or Walmart. Thus, Web Content Mining is mining statistics from the content material of net pages which consist of text, graphics, tables, facts blocks and statistics documents.

**Challenges in Web Content Mining:**

* Personalization of Web Space: Web space personalising is an important challenge in mining the Web contents. It requires a gadget which responds to person queries with the aid of doubtlessly aggregating data from numerous sources in a manner which is based on who the consumer is. So, the extraction of appropriate content is always challenge.
* Web Information integration and schema matching: In internet we can locate comparable facts differently. Identifying the semantically comparable records is once more trouble with content material mining.
* Extraction of Opinions: With help of opinions from online sources like forums, blogs, dialogue rooms is of first-rate significance for advertising intelligence.
* Knowledge Combination: Ideapyramids are very important, and it is a time-consuming activity if it is created manually. The predominant software is to arrange the portions of facts on the Web to provide the consumer a coherent image of the theme domain.
* Segmenting NetsidesRacket Detection: We needs the main content of the Webpage, but in some case, there is risk that it may also comprise noises, in the shape of advertisement, navigation hyperlinks, copyright notices. So routinely segmenting the webpage to extract the predominant contents is a fascinating problem.

**Web Content Mining Tools:**

* CanopyFighter
* ComputerizationWherever 6.1 (CW)
* Web Info Extractor (WIE)
* Mozenda
* Web Content Extractor

1. Web Structure Mining

Web Structure Mining (WSM) deals with mining information from the semi-structured nature of the Web. It is used to find the pattern behind the connectionassemblies of the network pages based on their hyperlink network topology. The main source of this mining are Hyperlinks and Document Structure. Web Structure Mining reduces some of the main problems of WWW.

1. **Faster Indexing:** As WSM deals with discovering the model underlying the Web hyperlink structure it becomes easier to index web pages even if the information provided is vast.
2. **Relevant Search Result:** WSM uses Page Ranking mechanism to enhance the search results. For this we need to understand two basic terms Firm Web Pages and Hubs

**Authoritative Web Pages and Hubs:**

If an author of a webpage makes a hyperlink that points additionalNetworkLeaf can be measured to an authorization of the other web leaf. When many authors points to that same page that results in the collective endorsement of a given page indicates the significance of that page. This leads to the discovery of authoritative web pages. For example, you can find several web pages pointing to [www.facebook.com](http://www.facebook.com) or [www.twitter.com](http://www.twitter.com) . These are the examples for authoritative web pages(Authorities).

A traditional of sides that delivers collection of associations to authoritative webpages are called Hubs. In general, good hub page for a topic point to numerous authoritative pages and a proper authority web page is pointed by using many right hub pages on the equal topic.

Hubs

Authorities

The concept of Authoritative Web Pages and Hubs were used in HITS (Hyperlink-Induced Topic Search) Procedure in listing the search outcome. This method was precursor to Page Rank concept.

**Page Rank:**

The concept of Page Rank was introduced by Google. It is also a link analysis technique similar to HITS. In Page Ranking a numerical value is assigned to each element in hyper-linked set of documents. It indicates the relative importance of the webpage over other. With the help of a Web Crawler called Googlebot google counts links and collect other information on webpages.

1. Challenges in Web Structure Mining

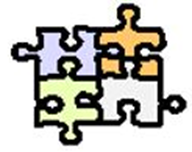
WSM also have the basic Web Mining Challenges. Moreover from its own perspective it has its own internal challenges that affect the actual mining process.

* **Google Bomb**: It is one of the classic examples of page rank manipulation. Google Bomb occurs when some group of people conspire to manipulate Google Search engine into returning a unique internet web page as the topmost end result via linking a unique phrase or phrase to the website.
* **Link Farming:** It is an automated activity where pages are linked without any relevance. So when we run into a web page there will be a random collection of links to other website without any relevance.
* **Extraction of Data from the Web**: Data in Web is semi structured. We cannot use SQL like query languages to handle these kinds of data. Hence data/information extraction is a challenging task.

1. Web Usage Mining

Web Usage Mining (WUM) is an application of Data Removal that deals with the detection of interesting formson or after users Web usage. It mainly focus on extracting records from server logs. Usage facts captures the beginning of customers and their behavior of looking at a website. Earlier this technique was used by marketing giants to improve profit but now search engines also have implemented to enhance user’s browsing experience thereby optimizing their search results. Other way to find Web user traffic is to sniff TCP/IP packets and plugin each to the web server.

**Web Usage Mining Process:**

****

User Models

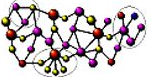
Web & Application Server Logs

Data pre-processing

User Transaction Database

Usage Mining

Pattern Analysis

**Pattern**

Let us see how the user is being identified.

* Embedding Session Id’s: For every user, whenever a request is made to the server a session is created upon successful request-response is made. This uses dynamically generated pages to associate ID with every hyperlink.
* Capturing IP Address: This is one of the oldest techniques used to log user’s activity. Requires no additional technology as every IP address is treated as a unique user. But the main drawback is now the IP is dynamic.
* Registration: Some websites uses explicit registration process to track the activity of the user which is an overhead.
* Cookies: It actually saves the ID on client machine which can track repeat visit from the same browser.
* Software Agents: This is typically a technique that is very effective but often rejected by the users as the program is directly loaded to the browsers and it sends back the usage details.

**Challenges in Web Usage Mining:**

As WUM is unity of the prominent field in Web Mining, it poses great number of challenges.

* Personalization of Web Content: This is one of the evolving applications of WUM. Still it has not reached the perfection. For instance, it should be able to anticipate in real time the current navigation pattern and previously extracted pattern from weblogs so that each user can feel the personalized browsing experience. The barriers to this include the proxy servers and Dynamic IP’s.
* User Identification: In order for WUM to be effective the basic level of details should be available. We have seen various user identification techniques and these techniques are having several implementation issues. For instance, use of Dynamic IP’s, embedding session id’s cannot capture the repeat visitors which is an overhead for the web pages. Even the usage of cookies can be turned off by the users we cannot rely cookie based technique in WUM. Also some explicit registrations to a site can be effective to track the user activity, it is rejected by most of the users as it requires more time and effort.
* Proxy Servers, Anonymizers, Caching: We may wonder why the use of the proxy and caches seems to be a challenge as we all seen them to enhance or improve the performance earlier. Unfortunately, in the field of Web Mining it is a curse to some extent. With the use of proxy Server the traffic to the main web server is reduced. It acts as an intermediary between client who seeks resources from web servers which includes files, connection, webpage etc. As the proxies facilitates access to the contents it causes anonymity. Anonymizers are similar to proxies but it is a tool that hides the activity on the internet. Here the user’s computer identifying information like IP addresses will be hidden.

References

1. <http://en.wikipedia.org/wiki/Web_mining>

2. **Data Mining: Concepts and Techniques**, 3rd Edition, Jiawei Han, Micheline   
Kamber, Jian Pei, 2011