International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-2 Issue-1, March 2013

Investigation and Analysis of Current Web Mining Techniques as well as Frameworks

Marjan Eshaghi, S.Z. Gawali

Abstract- Everyone using the Web, experiences how the connection to a popular web site may be very slow during rush hours and it is well known that web users tend to leave a site if the wait time for a page to be served exceeds a given value. Therefore, performance and service quality attributes have gained enormous relevance in service design and deployment. This has led to the development of Web benchmarking tools largely available in the market. One of the most common critics to this approach, is that synthetic workload produced by web stressing tools is far to be realistic. Moreover, Web sites need to be analyzed for discovering commercial rules and user profiles, and models must be extracted from log files and monitored data. This paper deals with a benchmarking methodology based on the integrated usage of web mining techniques and standard web monitoring and assessment tools.

Keywords: Web Mining, Pattern Extraction, Usage Mining, Preprocessing.

I. INTRODUCTION

In this world of data Technology, accessing Information is that the most frequent task. A day we've got to travel through much quite info that we'd like and what we have a tendency to do? Simply browse the online and therefore the desired info is with America on one click. Today, net is taking part in such a significant role in our standard of living that it's terribly tough to survive while not it. The globe Wide net (WWW) has influenced plenty to each users (visitors) further because the electronic computer homeowners. The online website homeowners area unit able to reach to any or all the targeted audience Internationally. They're receptive their client 24X7. On the opposite aspect guests also are availing those facilities. To mine the fascinating knowledge from these data processing techniques will be applied. However the online knowledge is unstructured or semi structured. Thus we are able to not apply the info mining techniques directly. Rather another discipline is evolved known as net mining which may be applied to net knowledge. Net mining is employed to get interest patterns which may be applied to several universe issues like up internet sites, higher understanding the visitor's behavior, product recommendation etc. Web usage mining could be a method of learning info from user a way to use internet sites. Web page mining could be a method of learning info from texts, pictures and different contents. Net structure mining could be a method of learning info from linkages of websites [11].

Revised Manuscript Received on 30 March 2013.
* Correspondence Author
Marjan Eshaghi, Information Technology Department, College of Engineering, Bharati Vidyapeeth University, Pune, India.
S.Z. Gawali, Information Technology Department, College of Engineering, Bharati Vidyapeeth University, Pune, India.

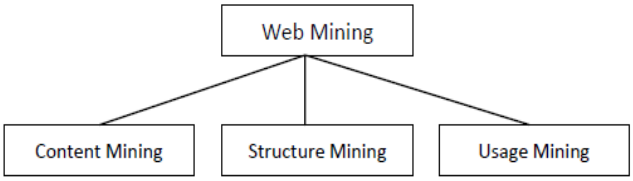

Figure 1: Web mining classification

TABLE 1
THE RELATIONSHIP AMONG THE DIFFERENT AREAS OF WEB MINING [11].

Type	Structure	Form	Object	Collection
Usage	Accessing	Click	Behavior	Logs
Content	Pages	Text	Index	Pages
Structure	Map	Hyperlinks	Map	Hyperlinks

These three approaches tries to extract information from net generate some helpful result from that information and apply the result to bound universe issues. Net Usage Mining is that the method of applying data processing techniques to the invention of usage patterns from information extracted from blog files.

Web usage mining is one among the distinguished analysis space attributable to these following reasons. a) One will keep track of antecedently accessed pages of a user. These pages are often accustomed establish the standard behavior of the user and to create prediction concerning desired pages. Therefore personalization for a user is often achieved through net usage mining. b) Frequent access behavior for the users are often accustomed establishes required links to enhance the general performance of future accesses. Perfecting and caching policies are often created on the idea of oft accessed pages to enhance latency time. c) Common access behaviors of the users are often accustomed improve the particular style of websites and for creating alternative modifications to an online website. d) Usage patterns are often used for business intelligence so as to enhance sales and ad by providing product recommendations.

Five major steps followed in net usage mining square measure [9]

1. Information assortment – blog files, that keeps track of visits of all the guests
2. Information Integration – Integrate multiple log files into a single file
3. Information preprocessing – cleanup and structuring information to organize for pattern extraction
4. Pattern extraction – Extracting fascinating patterns

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

5. Pattern analysis and mental image – Analyze the extracted pattern
6. Pattern applications – Apply the pattern in universe issues

II. LITERATURE REVIEW

Web Data

In internet Usage Mining, knowledge are often collected in server logs, browser logs, proxy logs, or obtained from AN organization\'s info. These knowledge collections disagree in terms of the placement of the information supply, the styles of knowledge obtainable, the phase of population from that the information was collected, and ways of implementation.

There are several styles of knowledge that may be employed in internet Mining [1].

1. Content: The visible knowledge within the web content or the data that was meant to be imparted to the users. a serious a part of it includes text and graphics (images).

2. Structure: knowledge that describes the organization of the web site. It's divided into varieties. Intra-page structure data includes the arrangement of varied markup language or XML tags at intervals a given page. The principal styles of inter-page structure data are the hyper-links used for web site navigation.

3. Usage: knowledge that describes the usage patterns of web content, like information processing addresses, page references, and therefore the date and time of accesses and varied different data counting on the log format.

Data Sources

The knowledge sources utilized in net Usage Mining could embrace net data repositories like [7]:

1. net Server Logs – These are logs that maintain a history of page requests. The W3C maintains a customary format for net server log files, however alternative proprietary formats exist. Newer entries are generally appended to the top of the file.

Information concerning the request, together with shopper information science address, request date/time, page requested, HTTP code, bytes served user agent and referrer are generally additional.

This knowledge will be combined into one file, or separated into distinct logs, like associate degree access log, error log, or referrer log. However, server logs generally don't collect user-specific info. These files are typically not accessible to general web users, solely to the webmaster or alternative body person. An applied mathematics analysis of the server log is also accustomed examine traffic patterns by time of day, day of week, referrer, or user agent. Economical computing machine administration, adequate hosting resources and also the fine calibration of sales efforts will be power-assisted by analysis of the online server logs. Promoting departments of any organization that owns a web site ought to be trained to grasp these powerful tools.

2. Proxy Server Logs - an internet proxy could be a caching mechanism that lies between shopper browsers and net servers. It helps to scale back the load time of web content furthermore because the network traffic loads at the server and shopper aspect. Proxy server logs contain the protocol requests from multiple purchasers to multiple net servers. This might function a knowledge supply to get the usage pattern of a bunch of anonymous users, sharing a typical proxy server.

3. Browser Logs – varied browsers like Mozilla . will be changed or varied JavaScript and Java applets will be accustomed collect shopper aspect knowledge. This implementation of client-side knowledge assortment needs user cooperation, either in practicality of the JavaScript and Java applets, or to voluntarily use the changed browser [8]. Client-side assortment scores over server-side assortment as a result of it reduce each the larva and session identification issues.

Information Obtained

1. Variety of Hits: This variety typically signifies the amount of times any resource is accessed in an exceedingly web site [5]. Successful could be a request to an internet server for a file (web page, image, JavaScript, Cascading sheet of paper, etc.). once an internet page is uploaded from a server the amount of "hits" or "page hits" is adequate to the amount of files requested. Therefore, one page load doesn't perpetually equal one hit as a result of usually pages are created of different pictures and different files that pull together the amount of hits counted.

2. Variety of Visitors: A "visitor" is strictly what it feels like. It is a human WHO navigates to your website and browses one or a lot of pages on your site.

3. Traveler Referring Website: The referring web site offers the data or address of the web site the web site that referred the actual website in thought.

4. Traveler Referral Website: The referral web site offers the data or address of the web site the web site that is being cited by the actual website in thought.

5. Time and Duration: This data within the server logs offer the time and period for a way long the web site was accessed by a selected user.

6. Path Analysis: Path analysis offers the analysis of the trail a selected user has followed in accessing contents of an internet site [5].

7. Traveler IP address: This data offers the net Protocol (I.P.) address of the guests WHO visited the web site in thought.

8. Browser Type: This data offers the data of the sort of browser that was used for accessing the web site.

9. Cookies: A message given to an internet browser by an internet server. The browser stores the message in an exceedingly document known as cookie. The message is then sent back to the server every time the browser requests a page from the server. The most purpose of cookies is to spot users and probably prepare tailor-made websites for them. Once you enter internet web site mistreatment cookies, you will be asked to fill out a type providing such data as your name and interests. This data is prepackaged into a cookie and sent to your browser that stores it for later use. Following time you visit a similar internet site, your browser can send the cookie to the online server. The server will use this data to gift you with custom websites. So, for instance, rather than seeing simply a generic welcome page you would possibly see a welcome page together with your name on that.

10. Platform: This data offers the sort of software package etc. that was wont to access the web site. Possible Actions

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

1. Shortening ways of High visit Pages: The pages that area unit of accessed by the users may be seen on follow a selected path. These pages may be enclosed in Associate simply accessible a part of the web site therefore leading to the decrease within the navigation path length.

2. Eliminating or Combining Low Visit Pages: The pages that don't seem to be oft accessed by users may be either removed or their content may be incorporate with pages with frequent access.

3. Redesigning Pages to assist User Navigation: to assist the user to navigate through the web site within the absolute best manner, the data obtained may be accustomed design the structure of the web site.

4. Redesigning Pages for program improvement: The content moreover as different info within the web site may be improved from analyzing user patterns and this info may be accustomed design pages for program Optimization so the search engines index the web site at a correct rank.

5. Facilitate Evaluating Effectiveness of Advertising Campaigns: vital and business vital advertisements may be place au fait pages that area unit of accessed.

Web Usage Mining Process:

The main processes in net Usage Mining are: Preprocessing: information preprocessing describes any style of process performed on data to arrange it for one more process procedure. Unremarkably used as a preliminary data processing follow, information preprocessing transforms the info into a format that may be a lot of simply and effectively processed for the aim of the user. The various sorts of preprocessing in net Usage Mining are [1][5][11]:

1. Usage Pre-Processing: Pre-Processing regarding Usage patterns of users.

2. Content Pre-Processing: Pre-Processing of content accessed.

3. Structure Pre-Processing: Pre-Processing associated with structure of the web site.

Pattern Discovery: net Usage mining will be accustomed uncover patterns in server logs however are usually dispensed solely on samples of knowledge. The mining method is ineffective if the samples aren't a decent illustration of the larger body of knowledge. The subsequent area unit the pattern discovery ways [9].

1. Applied math Analysis
2. Association Rules
3. Clustering
4. Classification
5. Ordered Patterns
6. Dependency Modeling

Pattern Analysis: this can be the ultimate step within the net Usage Mining method. When the preprocessing and pattern discovery, the obtained usage patterns area unit analyzed to filter uninteresting data and extract the helpful data. The ways like SQL (Structured question Language) process and OLAP (Online Analytical Processing) will be used [9].

Web Usage Mining Areas

1. Personalization
2. System Improvement
3. Website Modification
4. Business Intelligence
5. Usage Characterization

Web logs

A web server log file could be a straightforward plain document that records data whenever a user requests a resource from an internet website. This file is opened once the online services of a server starts and stay open because the server responds to user requests [7]. Diary files offer net directors with several helpful quite data like:

- That pages of your internet site were requested?
- What area unit the errors that folks encounter?
- What's the standing came back by the server upon user request?
- What number bytes sent from the server to the user?

Analyzing this information might reveal necessary patterns. Usually there are unit four sorts of server logs [8]:

1. Access log file
2. Error log file
3. Agent log file
4. Referrer log file

The first two types are the most commonly used. The agent and referrer logs may or may not be enabled at the server. Access log file contains data of all incoming requests and lets you track and get information about clients of the server. Error log file lists internal server errors. This information enables server administrators to correct site content or to detect anomalous activities. Agent log file provides information about user's browsers, operating system and browser version. Referrer log provides information about the link that redirects visitors to my site. Our work in this paper will focus on web access log files or simply web log files.

III. TECHNIQUES

A. data processing

There square measure two categories knowledge of knowledge of information mining particularly i) to summarize or characterize general properties of knowledge in repository that is named Descriptive and ii) to perform logical thinking on current data, to create predictions supported the historical information that is named Prescriptive. There square measure numerous data processing techniques on the market that can also be applied to internet data processing. Few techniques square measure listed below [11].

1) Association Rules Mining: once the book data processing ideas and Techniques is bought, four-hundredth of the time the book information System is bought along, and twenty fifth of the time the book information Warehouse is bought along. Those rules discovered from the dealing information of the book store may be wont to arrange the means of the way to place those connected books, which may more create those rules a lot of sturdy [11].

2) Serial Pattern Mining: Association rule mining doesn't take the time stamp under consideration; the rule may be get A=>Buy B. If we tend to take time stamp under consideration then we are able to get a lot of correct and helpful rules such as: get A implies get B at intervals per week, or sometimes individuals get A hebdomadally. As we are able to see with the second quite rules, business organizations will create a lot of correct and helpful prediction and consequently create a lot of sound choices.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Information consists of sequences of values or events that amendment with time is named time-series information [Han and Kamber 2000], a time-series information records the valid time of every dataset. for instance, in an exceedingly time-series information that records the sales dealing of a grocery, every dealing includes an additional attribute indicate once the dealing happened. Time series information is wide wont to store historical information in an exceedingly diversity of areas like, monetary information, medical information, scientific information so on. Completely different mining techniques are designed for mining time-series information, primarily there square measure four types of patterns we are able to get from numerous sorts of time series data:1) analysis, 2) Similarity search, 3) serial patterns and 4) Periodical patterns. Serial patterns: serial pattern mining is attempting to seek out the relationships between occurrences of serial events, to seek out if there exists any specific order of the occurrences. we are able to notice the serial patterns of specific individual things; additionally we are able to notice the serial patterns cross completely different items. Serial pattern mining is wide employed in analyzing of polymer sequence. AN example of serial patterns is that each time Microsoft stock drops five-hitter, IBM stock also will drops a minimum of wheel drives at intervals 3 days [9].

3) Classification: Classification is to create (automatically) a model that may a category of objects thus on predict the classification or missing attribute worth of future objects (whose class might not be known). It's a ballroom dancing method. Within the initial method, supported the gathering of coaching information set, a model is built to explain the characteristics of a collection of knowledge categories or ideas. Since information categories or ideas square measure predefined, this step is additionally called supervised learning (i.e., that category the coaching sample belongs to is provided). Within the second step, the model is employed to predict the categories of future objects or information. A call tree for the category of get laptop computer, indicate whether or not or not a client is probably going to get a laptop computer. Every internal node represents a call supported the worth of corresponding attribute, additionally every leaf node represents a category (the worth of get laptop=Yes or No). Once this model of get laptop computer has been designed, we are able to predict the probability of shopping for laptop computer supported a replacement customer\'s attributes like age, degree and profession. That info may be wont to target customers of bound merchandise or services, particularly wide employed in insurance and banking [5][11].

4) Clustering: Classification may be taken as supervised learning method, clump is another mining technique the same as classification. But clump may be an unattended learning method. Clump is that the method of grouping a collection of physical or abstract objects into categories of comparable objects, in order that objects at intervals identical cluster should be the same as some extent, additionally they must be dissimilar to those objects in different clusters. In that record belongs that class is predefined, whereas in clump there's no predefined categories. In clump, objects square measure sorted along supported their similarities. Similarities between objects square measure outlined by similarity functions, sometimes similarities square measure quantitatively specific as distance or different measures by corresponding domain specialists. For instance, supported the expense, deposit and draw patterns of the shoppers, a bank will clump the market into completely different teams of individuals. For various teams of market, the bank will give different totally completely different completely different types of loans for homes or cars with different budget plans. During this case the bank will give a more robust service, and additionally check that that each one the loans may be saved [11].

B. Log Files

In order to manage a web server effectively, it is necessary to get feedback about the activity and performance of the server as well as any problems that may be occurring. Web server creates and maintains log files for this purpose. A Web log is a file to which the Web server writes information each time a user requests a resource from that particular site [1].

1) Log Formats: W3C maintains a standard format for web server log files, but other proprietary formats exist. For example IIS provides six different log file formats which are used to track and analyze information about IIS-based sites and services such as 1. W3C Extended Log File Format, 2. W3C Centralized Logging, 3. NCSA Common Log File Format, 4. IIS Log File Format, 5. ODBC Logging, 6. Centralized Binary Logging. In addition to the six available formats, custom log file format can also be configured.

A log file in the W3C extended format contains a sequence of lines containing ASCII characters. Each line may contain either a directive or an entry. Entries consist of a sequence of fields relating to a single HTTP transaction. Fields are separated by white space. If a field is unused in a particular entry dash "-" marks the omitted field. Directives record information about the logging process itself. Lines beginning with the # character contain directives. The following directives are defined in the W3C Extended format: The following is an example of a record in the extended log format that was produced by the Microsoft Internet Information Server (IIS) [1]:

#Software: Microsoft Internet Information Server 4.0
#Version: 1.0
#Date: 1998-11-19 22:48:39
#Fields: date time c-ip cs-username s-ip Cs-method cs-Uri-stem cs-uri-query scstatus sc-bytes cs-bytes time-taken csversion cs(User-Agent) cs(Cookie) cs(Referrer)
1998-11-19 22:48:39 206.175.82.5 - 208.201.133.173 GET /global/images/navlineboards.gif - 200 540 324 157 HTTP/1.0 Mozilla/4.0+(compatible;+MSIE+4.01;+Windows+95) USERID=CustomerA;+IMPID=01234 http://yourturn.rollingstone.com/webx?98@@webx1.html

Description of headers

c Client
s Server
r Remote
cs Client to Server.
sc Server to Client.
sr Server to Remote Server, this prefix is used by proxies.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Rs Remote Server to Server, this prefix is used by proxies.
x Application specific identifier.
Apache web server maintains Common Log Format and Combined Log Format

Common Log Format.

Log Format "%h %l %u %t \"%r\" %>s %b"
common
122.163.111.210 - - [22/Oct/2010:04:15:03
-0400] "GET
/imagesnew/misc_arrow_animated.gif
HTTP/1.1" 404 494

Combined Log Format

Another commonly used format string is called the Combined Log Format. It can be used as follows.
LogFormat "%h %l %u %t \"%r\" %>s %b
\"%{Referer}i\" \"%{User-agent}i\""
Combined
This format is exactly the same as the Common Log Format, with the addition of two more fields. Each of the additional fields uses the percent-directive %{header}i, where header can be any HTTP request header. The access log under this format will look like:
127.0.0.1 - frank [10/Oct/2000:13:55:36 -
0700] "GET /apache_pb.gif HTTP/1.0" 200
2326 "http://www.example.com/start.html"
"Mozilla/4.08 [en] (Win98; I ;Nav)"
The additional fields are:
"http://www.silicon.ac.in/sitsbp/index.ht
ml"
(\"%{Referer}i\")
The "Referer" gives the site that the client reports having been
referred from.
"Mozilla/4.0 (compatible; MSIE 6.0;
Windows NT 5.1; GTB0.0; SV1; .NET CLR
2.0.50727; .NET CLR 3.0.04506.30; .NET
CLR 3.0.4506.2152; .NET CLR 3.5.30729;
RediffIE8)"
(\"%{User-agent}i\")
The User-Agent HTTP request header. This is the identifying information that the client browser reports about itself. More recent entries are appended to the end of the file. These data can be stored into a single file, or separated into distinct logs, such as an access log, error log, or referrer log. Web usage mining research focuses on finding patterns of navigational behavior from users visiting website. These patterns of navigational behavior can be valuable when searching answers to questions like: How efficient is our website in delivering information? How the users perceive the structure of the website? Can we predict user's next visit? Can we make our site meeting user needs? Can we increase user satisfaction? Can we target specific groups of users and make web content personalized to them? Answer to these questions may come from the analysis of the data from log files stored in web servers. Web usage mining has then become a necessary task in order to provide web administrators with meaningful information about users and usage patterns for improving quality of web information and service performance. Successful websites may be those that are customized to meet user preferences both in the presentation of information and in relevance of the content that best fits the user.

Drawbacks

It may incur some extra overhead particularly once the Java applications program is loaded for the primary time.
• Java scripts, on the opposite hand, consume very little interpretation time however cannot capture all user clicks (such as reload or back buttons). These strategies can collect solely single-user, single-site browsing behavior.
• A changed browser is far a lot of versatile and can permit knowledge assortment a couple of single user over multiple Websites. the foremost tough a part of exploitation this methodology is convincing the users to use the browser for his or her daily browsing activities.

Privacy Issues:

Personal privacy has forever been a significant concern during this country. In recent years, with the widespread use of net, the issues regarding privacy have increase staggeringly. Attributable to the privacy problems, some folks don't look on net. They're afraid that someone could have access to their personal info then use that info in Associate in unethical way; so inflicting they damage.
Though it's against the law to sell or trade personal info between completely different organizations, merchandising personal info have occurred. For instance, in step with laundry Post, in 1998, CVS had sold their patient's prescription purchases to a unique company.
Security issues: Though firms have lots of private info regarding United States of America obtainable on-line, they are doing not have spare security systems in situ to shield that info. For instance, recently the Ford Motor credit company had to tell 13000 of the customers that their personal info as well as Social Security variety, address, account variety and payment history were accessed by hackers WHO stone-broke into info happiness to the Experian credit news agency [6].
This incidence illustrated that firms square measure willing to disclose and share your personal info, however they\'re not taking care of the data properly. With such a lot personal info obtainable, fraud may become a true drawback.

IV. CONCLUSION

As we tend to mentioned higher than, during this project we tend to bestowed new approach in preprocessing of journal files for internet intrusion detection. We tend to mention the various steps during this method and therefore the variations in these steps from internet usage mining. Additionally, we tend to illustrate a way to mix log files with completely different formats in one commonplace format exploitation XML. We tend to provided algorithms to mix those log files. These algorithms are enforced exploitation c# code and underneath windows panorama software system.

REFERENCES

1. "XML Based Web Usage Mining In Server Logs", Y.S.S.R Murthy, L.Balaji & Lakshmi Tulasi.Ambat.
2. A. Hamami, M. Ala'a, S. Hasan. (2006). Applying Data Mining Techniques in Intrusion Detection System on Web and Analysis of Web Usage, Information Technology Journal, 2006.
3. C.J. Ezeife, J. Dong, A.K. Aggarwal. (2007). SensorWebIDS: A Web Mining Intrusion Detection System, International Journal of Web Information Systems, volume 4, pp. 97-120, 2007.
4. C. Kruegel, G. Vigna. (2003). Anomaly Detection of Web-based Attacks, CCS, 2003.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

5. G. Shiva, N.V. Suba, U. Dinesh. (2010). Knowledge Discovery from Web Usage Data: A survey of Web Usage Pre-processing Techniques, Springer, 2010.
6. Andrews, M.: Guest Editor's Introduction: The State of Web Security. IEEE Security and Privacy, 4, 4, 14--15 (2006)
7. K.R. Suneetha, Dr. R. Krihnamoorthi. (2009). Identifying User Behavior by Analyzing Web Server Access Log File, IJCSNS, 2009.
8. L.K. Joshila Grace, V.Maheswari, Dhinaharan Nagamalai. (2011). Analysis of web logs and web user in web mining, IJNSA, 2011.
9. Jaideep Srivastava , Robert Cooley, Mukund Deshpande, Pang-Ning Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations, Volume 1, Issue 2- Pages 12-23.
10. Adel T. Rahmani and B. Hoda Helmi, EIN-WUM an AIS-based Algorithm for Web Usage Mining, Proceedings of GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA, ACM978-1-60558-130-9/08/07 (Pages 291-292)
11. Ms. Sarika Y. Pabalkar Pad Dr. D.Y, Web Text Mining for news by Classification. International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 6, August 2012

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)