THOMSON REUTERS ENDNOTE

[Sehnaz*, 5(5): May, 2016]

ISSN: 2277-9655
Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

MINING OF WEB LOG FILES USING RELEVANT COMPUTING TECHNIQUES FOR IMPROVING FUTURE ANTICIPATION USAGE OF WEB NAVIGATION

N.Sanfia Sehnaz*, Dr.I.Elizabeth Shanthi

* Research scholar Computer Science, Avinashilingam Institute for Home Science and Higher Education for women university, India.

Associate professor in Computer Science, Avinashilingam Institution for Home Science and Higher Education for Women University, India.

ABSTRACT

The Internet has evolved extensively over the past few decades. Web navigation refers to the process of navigating a network of information resources in the World Wide Web, which is organized as hypertext or hypermedia. The navigation related to web navigation usability gets solved by comparing the actual and anticipated usage patterns. The actual usage pattern removed from web server logs are sporadically recorded in operational websites for handling the log data. This process is used to identify the users, user session and user task oriented transactions. The pattern can be discovered among the actual usage path by using the algorithms of data mining generally the ideal user's interactive path models are framed by cognitive experts based on the cognition of user behavior, which is utilized to pull out the anticipated usage, that includes information about both the time required for user-oriented tasks and the mechanism to identify the user navigation problems here the usability issues get detected from the deviation of the data. It is observed that Genetic algorithms can be used as optimization methods and for corrective action to improve the web navigation usability.

KEYWORDS: Data mining, Log files, Cognition models, Genetic Algorithms.

INTRODUCTION

In the modern digital world easy-to-use of web system and relevant information, retrieval form. World Wide Web is becoming a mandatory requirement for every business web usage means effective, efficient and satisfy able retrieval of data form web. This can be achieved by considering the web design principle called structural firmness, functional convenience and presentational delight. Structural firmness means the website is secured for transaction next relates the ease of navigation through web and ease of user's interaction. Presentational delight makes the users interface more adorable. In order to retrieve useful information from very large databases. Various strategies that are part of data mining area are used.

Data mining is a multidisciplinary computer science field. That has reckoning process of discovering patterns in large data sets using interaction of the database systems, artificial intelligence and machine learning statistics. The main goal of the data mining process is to haul out information from a data set and transform it into an understandable structure for further use. Data mining techniques are used in many research areas such as mathematics, cybernetics, genetics and marketing. Web mining, a type of data mining used in customer relationship management, takes advantage of the huge amount of information gathered by a website to look for patterns in user behavior. The recent years have seen the flourishing of research in the area of web mining and specifically of web usage mining. Web usage mining is that area of web mining which pulls out the exceptional knowledge from logging information produced by web servers. Web usage mining is that part of web mining which deals with mining of the knowledge from server log files; source data consist of textual logs. Web content mining is that part of web mining which rivets on the raw information available in web pages. Web Structure mining is that parts of web mining which bring together on the structure of web sites (e.g., links to every other page of the sites). Web usage mining is that part of web mining which deals with mining of the knowledge from server log files; source data consist of textual logs.

ISSN: 2277-9655
Impact Factor: 3.785

Genetic Algorithm (GA) is a branch of artificial intelligence which was inspired by Darwin's theory of living organisms in which successful organisms were produced as a result of evaluation. GA is a search algorithm based on natural selection and natural genetics. The main implication of genetic algorithm is the continued existence of the fittest which is also known as natural selection. The genetic algorithm uses a structured population model in which each genome's reproductive pattern is selected from within its local neighborhood. This genetic algorithm lets improve the web usage navigations pattern of the web mining concept.

RELATED WORK

Logs, web usage and usability

Web log is a website that contains a series of entries arranged in reverse chronological order, often updated on continuously with new information about particular field. The information can be written by the site owner, brought together from other Web sites or other sources, or contributed by users [4][5]. Two types of logs such as server side logs and client side logs are commonly used for web usage and usability analysis. Server-side logs can be automatically originated by web servers, by every request of the users.

By bringing together these logs, web workload was characterized and used to suggest the performance improvement for the internet web server. Due to the unstable web traffic, massive user population and diverse usage environment, coverage-based testing is insufficient to improve the quality of web application. The server-side logs have been used to paradigm the model of the web usage for the usage-based web testing on the other hand client side logs can grab exact usage data for usability analysis, because they allow low-level user interaction events such as keystroke and mouse movements that to be recorded [2][7]. The properties and benefits of the log files are listed below [11]:

a) Automatic logging --→ No setup required, cost effective
b) Logging standard ---> Ubiquitous data
c) Server-side logging ---> No negative impact
d) IP address --→ Geographic location, identifying user
e) Standard formats --→ Easy use of data
f) Username --→ Identifying user
g) Timestamp --→ Knowing task time
h) Request --→ what users are doing
i) Referrer --→ How users arrived

Cognitive user models

In present days, there is popular growth to incorporate insights from cognitive science about the mechanisms, strengths and limits of human perception and cognition to understand the human factors involved in user interface design. Cognitive models mainly include GOMS, EPIC and ACT-R [2]. GOMS model consist of operators, selection rule, methods and goals. GOMS describes behavior and defines a static sequence of human in the high-level architecture [8]. As the low level cognitive architecture the Executive – Process/Interactive Control (EPIC) and Adaptive Control of Thought-Rational (ACT-R) these both models can be used for the implementation of high-level architecture. The important feature of low-level cognitive architecture is that they are equipped as computer programming system so that models may be specified, executed and compared with human performances [9]. Whereas the ACT-R model provides detailed process models of human performances with many complex task and helps the researches to specify the cognitive factors by developing cognitive models. ACT-R is mainly used to understand the decisions of the users for many links which satisfies the goals of the users.

ARCHITECTURE

The architecture of identifying the various usability processes is given in Fig 1 [2] that compares web usage pattern removed from server logs against predicted usage represented in some cognitive user models.

Web log files

Web log files contain the information of HTTP details of requested site. Let's see some of the sample log file and its working for the implementation phase. Data set which is used for the implementation phase contains the information

about the host details. The timestamp is in the format "DAY MON DD HH:MM:SS YYYY", where DAY is the day of the week, MON is the name of the month, DD is the day of the month, HH:MM:SS is the time of day using a 24-hour clock, and YYYY is the year.

The three different modules identified are:

- Usage Pattern Extraction
- IUIP Modeling
- Usability problem identification

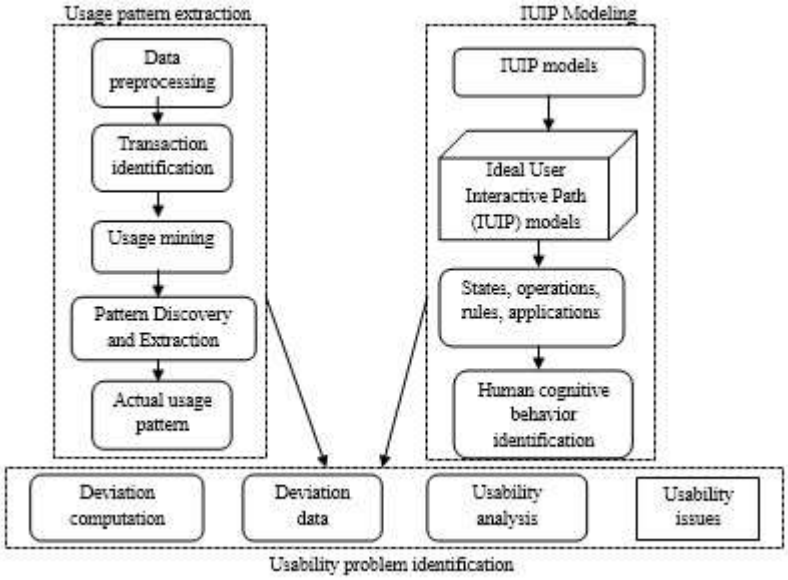

Fig 1. Architecture for identifying usability problems

3.2 MODULES

3.2.1 Usage pattern extraction

The actual user behavior is extracted from the web server logs and by using the cognitive user models the anticipated user behavior gets captured, then between these two a comparison is performed. This deviation analysis would help us identify some navigation related usability problems.

Web server logs are data source. Each entry in a log contains the IP address of the commenced host, the timestamp, the requested Web page, the referrer, the user agent and other data [1]. Generally, the raw data need to be preprocessed and transformed into user sessions and transactions to haul out usage patterns.

Data preparation and preprocessing

Four domain-dependent tasks are included in data preparation and preprocessing [10]:

- Data Cleaning
- User Identification
- User session Identification
- Path completion

In data cleaning task, the files which are not important for the analysis get removed. The user identification gets collected from combination of IP, user agent, and referrer fields to get authentication. User session identification task, checks the single visit to the site gets stored as an activity record which get segmented as sessions. The elapse time and threshold points are established to get effective session records. In path completion task, the missed references in cache which are getting obtained from the knowledge of site topology and referrer information, along with temporal information from server logs.

Transaction identification
A task gets separated in to individual data as groups needed for web transactions. The knowledge of site topology and referrer information, along with temporal information obtained from server logs. A transaction differs from a user session in that the size of a transaction can range from a single page to all the visited pages in a user session. The pages get identified from the click stream based on the event models [12].

Trail tree construction
The collection of path get identified from the transaction occur in each sessions. The tire algorithm is used to construct a tree structure that also traps user visit frequencies, which is called a *trail tree*. In a trail tree, a complete path from the root to a leaf node is called a *trail*. Each node corresponds to the occurrence of a specific page in a transaction. It is annotated with the number of users having reached the node across the same trail prefix. The leaf nodes of the trail tree are also annotated with the trail names. The mining can be performed again based on some interesting category [12].

3.3 Ideal user interactive path model construction
The user behavior pattern gets traced by sequence of states and transmissions. The sequence of related operations rule gets specified for a series of transaction for a particular goal. IUIP model specifies both the path and the benchmark interactive time (no more than a maximum time) for some specific states (pages). The benchmark time can first be specified based on general rules for common types of web pages. IUIP models established by cognitive experts and designers specify the anticipated user behavior. Both the paths and time for user-oriented tasks anticipated by Web designers from the perspective of human behavior cognition are captured in this model [12].

3.4 Usability problem identification
The actual users' navigation trails we extracted from the aggregated trail tree are analyzed against corresponding IUIP models automatically. This comparison will yield a set of deviations between the two. Based on logical choices made and time spent by users at each page, the calculation of deviations between actual users' usage patterns and IUIP can be divided into two parts: local deviation calculation and temporal deviation calculation [12].

RELEVANT COMPUTING TECHNIQUES

Soft computing (SC) solutions are changeable and uncertain between 0 and 1. It has been a formal area of study in Computer Science in the early 1990s. Earlier computational approaches could model and specifically analyze only relatively simple systems. It should be pointed out that easiness and difficulty of systems are relative, and many conventional mathematical models have been hard-hitting and very productive. Soft computing deals with imprecision, uncertainty, partial truth, and approximation to achieve practicability, robustness and low solution cost. As such it forms the basis of a substantial amount of machine learning techniques. Recent trends tend to include evolutionary and swarm intelligence based algorithms and bio-inspired computation.

The principal constituents of Soft Computing (SC) are

- Fuzzy Logic (FL)
- Neural Computing (NC)
- Evolutionary Computation (EC)
- Machine Learning (ML)
- Probabilistic Reasoning (PR)

4.1 Evolutionary Computation
In computer science, evolutionary computation is a subfield of artificial intelligence (more particularly computational intelligence) that can be defined by the type of algorithms it is concerned with. These algorithms, called evolutionary algorithms, are based on adopting Darwinian principles. Evolutionary computing techniques mostly involve met heuristic optimization algorithms includes:

4.1.1 Genetic algorithm

Genetic algorithm is started with a set of solutions called population. Solution from one population are taken and used to form a new population. This is motivated by a hope, that the new population will be better than the old one. Solutions which are selected to form new solutions (offspring) are selected according to their fitness – the more suitable they are the more chances they have to reproduce. In genetic algorithm, the process of initialization, selection, crossover and mutation is repeated until either the average fitness (quality, probability, etc.) of the population has reached an acceptable value or one of the members (the star) has such incredible quality. There is a repair operator, which can generate a feasible solution from an infeasible solution. Many performances can be done for such repairing. A simple implementation is to randomly remove a selected customer from the solution until the solution is feasible. The anticipated usage of web navigation gets as input in order to predict in detail about the effective prediction of anticipated web navigation. This is repeated until some condition (improvement of the best solution) is satisfied.

4.1.2 Applying of Genetic algorithm

The information about the various usages of website by different users is present in the weblog. The input about the weblog for every single user is provided as an input by binary encapsulation. Those can be taken as input dataset for genetic algorithm process. The initial population is taken or considered by the single user data usage. The chromosome taken by considering the pair of address obtained from the user's weblog, in those addresses the cross over operation gets computed. The random selection of any value in the computed result had undergone with the mutation process. The newly obtained result as child get compared with parent value to decide the fitness of the result, which decided by the strength of the result obtained. The new offspring get developed either as the parent or the child depend on the result obtained from the strength. The fitness value used to predict the best population.

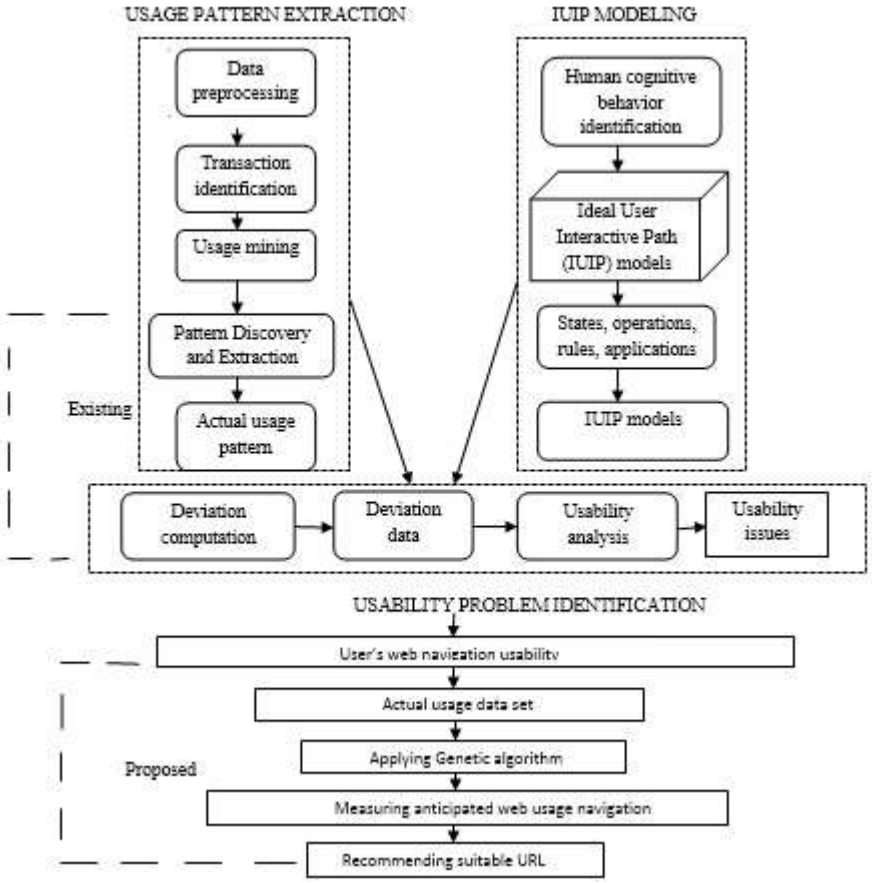

Fig 2: Proposed model

The fig 2 proposed processes undergoes number of iterations in order to obtain the best result. From the process the following observation are mentioned:

- The iterative process helps to collect the effective details about anticipated usage pattern from high granularity of data.
- The delay and frustration in finding the pattern get reduced and solve the usability problem and also supports the enhancement.
- The success rate for a task gets improved by improving the efficiency and time on task get measured using the usability.
- The scalability and effectiveness of a system get improved.

CONCLUSION

This paper analyses the impact of applying the computing techniques on web log files. A new method for the identification and improvement of navigation related web usability problems by checking extracted usage pattern against cognitive user model is suggested. The recent research work for predicting web navigation relies on using Usage Pattern Extraction, IUIP modeling and Usability Problem Identification. Meanwhile cognitive model like GOMS, ACT-R and EPIC have been successfully tested for better results. The selection and prediction of the feature set is done using cognitive model which has valuable impact on the performance of web navigation. The usability improvement in successive iteration can be obtained by progressively applying optimization techniques. It can be predicted that the usage of web navigation can be improved substantially by incorporating the relevant computing techniques.

REFERENCES

[1] T.carta, F.Paterno, and V.F.D. Santana, "Web usability probe: A tool for supporting remote usability evaluation of the web sites," in human-computer interaction –INTERACT 2011. New York, NY, USA: Springer, 2011.
[2] RuiliGeng& Jeff Tian members of IEEE have proposed a transaction paper on "Improving web navigation Usability by comparing Actual & Anticipated Usage".
[3] Anderson, j. R. (1983)."The architecture of cognition". Mahwah, nj, usa: lawrenceerlbaum.
[4] Federico Michele Facca, Pier Luca Lanzi, "Mining interesting knowledge from weblogs: a survey"Data & Knowledge Engineering 53 Elsevier (2005) 225–241.
[5] Haibin Liu, VladoKeselj "Combined mining of Web server logs and web contents for classifying user navigation patterns and predictingusers' future requests"Data & Knowledge Engineering 1 Elsevier (2007) 304–330 .
[6] M.F. Arlitt and C.L. Williamson, "Internet web servers:workload characterization and performance implication," IEEE/ACM trans.netw., vol.5, no.5, 631-645, oct 1997.
[7] G.Christou, F.E.Ritter , and R.J. Jacob, "CODEIN- A new notation for GOMS to handle evaluation of reality-based interaction style interfaces." Int.j. Human-Comput. Interactio, vol.28, no.3, pp.189-201, 2012.
[8] D.E.Kieras and D.E.Meyer . " An overview of the EPIC architecture for cognition and performance with application to human-computer interaction." Human comput.Interaction, vol. 12, no.4, pp.391-437, 1997.
[9] R.Cooley, B.Mobasher, and J.Srivastava., "Data preparation for mining world wide web browsing patterns," Know.Inf. Syst., vol. 1, no. 1, pp.5-32, 1999.
[10] http://vwhci.avestia.com/2014/PDF/006.pdf.
[11] http://www.ijstm.com/images/short_pdf/1452760169_178I.pdf