International Journal of Latest Trends in Engineering and Technology
pp. 170-174
DOI: http://dx.doi.org/10.21172/1.IRES.23
e-ISSN:2278-621X

WEB USER BEHAVIOR MODELING DISCOVERING BEHAVIOR PATTERNS AND ANALYSIS ON PATTERN ACCURACY

K. Abhirami[1] & M.K. Kavitha Devi[2]

Abstract-Web personalization systems are designed to improve user experiences by offering tailor-made services like web recommendations, adaptive web sites, web-enabled learning platforms, personalized web search etc. based on user interests and preferences recorded at log files of web, proxy servers or browsers. Personalization has varied applications through fittest solution meeting the model complexity and prediction accuracy. In this paper, we propose a framework to discover meaningful user browsing pattern on the web and convert the identified pattern into knowledge involving the web site being analyzed. Supervised machine learning techniques are used for discovering patterns of potential interests. Both single-valued, Bayesian inference & two-valued representations (belief & disbelief pair) Dempster-Shafer is proposed for inference and comparative precision identification. Proposed model thus addresses identification of browsing pattern that are known or expected in prior as well new pattern formation. Also, pattern analysis technique aims at identifying comparative performance of two approaches of pattern analysis.
Keywords: Machine learning, pattern discovery, pattern analysis, recommender systems.

1. INTRODUCTION

Discovery of web user profile and subsequent analysis plays a prime part in designing web based application areas such as healthcare, electronic-business, finance, marketing, web-based educational systems with personalized support, security including access to world wide web and social networking. To offer varied variety of user needs, personalized services are increasingly demanded. Catering to individual user needs is a challenging tasks and thus requires usage pattern discovery, followed by pattern analysis to support tailor-made services.

Deploying machine-learning and mining-techniques, dynamic web behavior based profile identification employing behavioral logging helps in extraction of useful patterns[6]. This makes more attractive user profile construction and is widely used in applications such as recommender systems. For example, mining the transaction history and ratings of ordered items, profile can be constructed to improve access and recommend items [7]. To assist and mediate impaired people by doctors and caretakers, profile based on user behavior is very important in healthcare industries. Similarly, in smart homes to assist people with specific needs, user behavior identification is very essential. E-Commerce organizations rely on websites to attract new customers and to retain existing. Personalized e-learning environments with user centered design methods through recommendations in learning management systems caters to ICT enabled learning. To achieve this web log files can be used to record user access patterns[2]. Thus, behavioral user profile discovered from activity situations, caters to customized recommendations[3].

2. DEFINITION AND BACKGROUND

2.1. Web usage data preprocessing

Multiple web server and application servers serve user requested content on the web. Integration of log files from several application and web servers is referred as Data fusion. Synchronization among web and application servers is required for this. Additional information that may not be essential for the purpose of data analysis is removed in the first step in web usage mining, which is Data cleaning. Chain of logged activities belonging to the same user which is activity record, identifies multiple sessions for each user. Section of user behavior record on the web, each denoting to a single visit to the site is Sessionization. Next possible and vital pre-processing task regularly carried out after sessionization is path completion. Set of n pageviews is the result of usage data pre-processing, denoted as P = {p1, p2, ⋯, pn}, and set of m user transactions, T = {t1,t2,⋯,tm}, where each ti in T is a subset of P.

2.2. Web usage pattern discovery

User guided modeling and dynamic user modeling techniques helps in discovering significant usage patterns. Former depending on individual information as given by the user, later depending on navigation behavior of users. Users with alike browsing behavior can be grouped using Clustering techniques. Similarity metric in the algorithm helps in identification of specific user profile based on the user performed activities[4].

[1] Department of Computer Science and Engineering, Kings College of Engineering, Thanjavur, Tamil Nadu, India.
[2] Department of Computer Science and Engineering, Thiagarajar College of Engineering, Madurai, Tamil Nadu, India.

In the perspective of Web Usage Mining, we can distinguish two cases of clusters which is user clusters and page clusters. Grouping pages having similar content is Web page clustering. User clustering is performed by grouping users by their similarity in navigational behavior. Similarity in visiting pattern at the same time period, changing user priorities for a page with varying time factor results in varied user clusters supporting the profiling process.
Each web page P distinctly symbolized by its linked URL accessed by the user is represented as Pg={pg1,pg2,…,pgm}. User access session is denoted as Sn={sn1,sn2,…..,snn} where each sni ϵSn is a subset of Pg. Session clustering is performed based on frequency of web page visit count, duration spent on a page, degree of user interest on a web page in a session. Clustering is an unsupervised approach to usage pattern discovery without predetermined labels.
Grouping data items into multiple identified classes is the role of Classification. Succinct model of the division of class labels in terms of predictor traits is the goal of this supervised learning technique. Testing instances are assigned class labels using the model.

2.3. Pattern analysis
Decisive objective of data mining task is to provide the analyst with insight into a particular domain. It comes from inspecting the discovered patterns and convert it to knowledge. A goal of the analysis drives the choice of what interestingness measures will be appropriate. A e-commerce site analyst examines the most popular patterns to gain insight into what is driving sales on the site. Support and confidence indentify the rules that are the most common and have the high correlation strength.

3. PROPOSED FRAMEWORK

Varied approaches and techniques are available to support effective recommendations for the users on the web. This becomes interesting due to the application of the fittest solution with minimal price. Role accomplishment of the recommender systems of a specific domain with the factor of performance improvement helps the web designers with the application of single or multiple techniques[1].

3.1. Pattern discovery
Supervised machine learning technique, classification is used to discover patterns. Decision forest algorithm is an ensemble learning method for classification. Several decision trees are built and then most accepted output class is chosen by the algorithm. Voting is a form of aggregation, in which every tree in a classification decision forest identifies a non-normalized frequency histogram of labels. Probability for every label is identified by aggregation process that sums histograms and normalizes the results. Trees with greater prediction confidence value will be assigned high weigh in the end decision of the ensemble.
For resampling, Bagging also called bootstrap aggregating can be used. In this method, using random sampling of actual dataset, each tree is grown with replacement until the size of the actual matches the dataset. If replicate method is chosen, each tree is trained on exactly the same input data. The determination of which split predicate is used for each tree node remains random and the trees will be assorted.

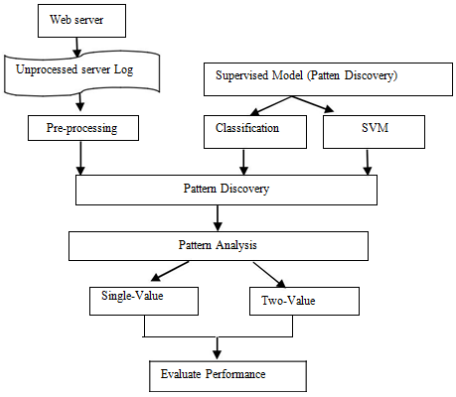


Fig. 1: Proposed system framework

Supervised learning model to analyze data and recognize pattern is Support vector machines (SVMs), used for classification and regression tasks. Prediction of two likely outcomes depending on continuous or categorical predictor variables is created by this classifier module. Using the set of training examples trained to belong to one of the two classes, SVM algorithms assigns new examples into the relevant category. The illustration is represented as points in space and is mapped. Examples of different categories are divided by wider gaps that is wide as possible. New samples are then mapped into relevant space as per prediction of category and to which side of the gap they fall on.

The feature space that contains the training examples is sometimes called a hyperplane, and it may have many dimensions. Support vector machines are among the earliest of machine learning algorithms, and SVM models have been used in many applications, from information retrieval to text and image classification.

3.2. Pattern analysis

Pattern analysis is the step that is designed to convert discovered rules, patterns into knowledge involving the website being analyzed. Identification of precise knowledge that is expected for better services is the major difficulty in pattern analysis process. Separation of the interesting results from those that are not particularly useful is one of the biggest challenges in pattern analysis.

3.3 Bayesian inference

For the Bayesian system a probability between the values of zero and one can be assigned to any belief. By definition, the probability against a belief, B, is equal to the probability not assigned in favor of the belief

P(B)=x,

P (¬B)=1-x

0<=x<=1

When new evidence, e is obtained about a belief, the probability is updated using P(B |e) = (P(e |B) P(B)) / P(e)

The new probability is referred to as being conditioned by the evidence, e. Both prior probabilities for all beliefs and probability of occurrence of particular piece of evidence is required in order to make use of Bayesian methods. Default value like 0.5 can be used in case of unknown belief value in advance.

3.4 Dempster-shafer inference

Dempster-Shafer (DS) method use a two-valued measure to describe the degree of belief or against a given statement to derive the inference. Evidence collected in favor and against b can be used for form support pair, [se, sl], where

se = essential support for

b sl = likely support for b

(1 – sl)= essential support for¬b

(1 – se)= likely support for¬b

(sl – se) = uncertainty of b

As an example, Evidence that belief b that web pages A,C are related

If all the evidence is in support of b, the DS pair is [1,1]

If all the evidence is against b, then DS pair is [0,0]

If data leads to 25% degree of belief that b is true and 40% degree of belief that b is false then [0.25, 0.6]

Degree of uncertainty is 35%

If no evidence pertaining to b, then DS pair is [0,1] giving uncertainty 100%

The values of Sn and Sp must satisfy the constraints:

Se + (1- Sl) <= 1

Se >=0, Sl >=0

Table 1: Page View Frequent Itemsets

Page view	Essential Support	Likely support
Home page, Courses, CSE	0.6	0.8
Home page, Scholarships	0.4	0.9
Home page, Resources	0.5	0.7
Home page, Courses, Student support	0.4	0.8
Home page, Courses, Students support, Skills	0.5	0.6
Home page, Links, e-books	0.1	0.8
Home page, sports	0.3	0.7

Table 2: Negative Necessary, Possible Support and Uncertainty for B

	Necessary Support for ¬B	Possible Support for¬ B	Uncertainty of B
Home page, Courses, CSE	0.2	0.4	0.2
Home page, Scholarships	0.1	0.6	0.5
Home page, Resources	0.3	0.5	0.2
Home page, Courses, Student support	0.2	0.6	0.4
Home page, Courses, Page view	0.4	0.5	0.1
Home page, Links, e-books	0.2	0.9	0.7
Home page, sports	0.3	0.7	0.4

Independent of the type of the source for generating an DS pair, pairs can be combined per Dempster's rule of combination to obtain single DS pair per belief.

4. EXPERIMENTAL RESULTS

In this section, we empirically evaluate the proposed approach and identify its performance. All experiments were performed on EDGAR Log File Data set. The EDGAR Log File Data Set contains information in CSV format extracted from Apache log files that record and store user access statistics for the SEC.gov website with the fields listed below:

ip: This variable provides the first three octets of the IP address
Date : Apache log file date
Time : Apache log file time
Zone : Apache log file zone
Cik : SEC Central Index Key (CIK) associated with the document requested
accession: SEC document accession number associated with the document requested
doc : This variable provides the filename of the file requested including the document extension
code : Apache log file status code for the request
filesize : document file size
idx : takes on a value of 1 if the requester landed on the index page of a set of documents
norefer : takes on a value of one if the Apache log file referrer field is empty, and zero otherwise
noagent : takes on a value of one if the Apache log file user agent field is empty, and zero otherwise
find : numeric values from 0 to 10,
crawler : web crawler
browser : web browser

To evaluate the performance of the system, with the initial steps of preprocessing, pattern discovery is made through modeling the system using Multiclass decision forest, classification technique with resampling methods of Bagging and Replicate. Accuracy results is found increasing with replicate resampling method.

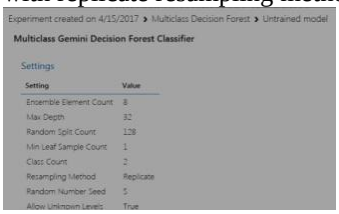


Fig. 2: Multiclass decision forest-resampling method: replicate

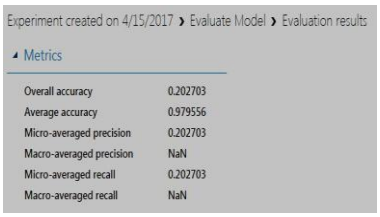


Fig. 3: Multiclass decision forest-resampling method: bagging

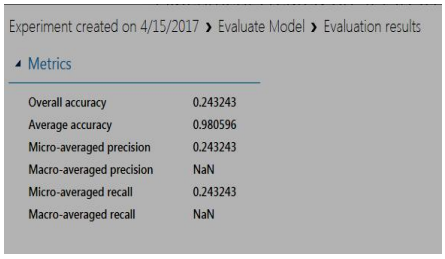


Fig. 4: Multiclass decision forest-resampling method: replicate

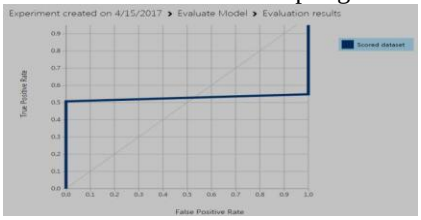


Fig. 5(a): Evaluation results–TP VS FP rate

To evaluate the effectiveness of the proposed system, accuracy is the parameter as shown in table below

Table 3: Accuracy Identification Parameter

	Items recommended by the system	Items not recommended by the system
Expected Item	True Positive(TP)	False Negative(FN)
Not an expected Item	False Positive(FP)	True Negative (TN)

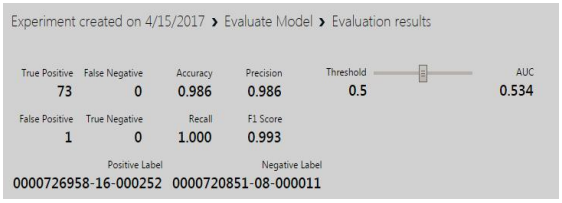
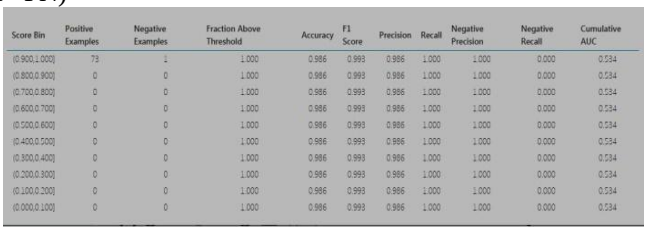


Fig. 5(b): Evaluation results

Recall = True Positive / (True positive + false negative)
Precision = True Positive / (True positive +False positive)
Accuracy =TP +TN /(TP+TN+FP+FN)



Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	73	1	1.000	0.986	0.993	0.986	1.000	1.000	0.000	0.534
(0.800,0.900]	0	0	1.000	0.986	0.993	0.986	1.000	1.000	0.000	0.534
(0.700,0.800]	0	0	1.000	0.986	0.993	0.986	1.000	1.000	0.000	0.534
(0.600,0.700]	0	0	1.000	0.986	0.993	0.986	1.000	1.000	0.000	0.534
(0.500,0.600]	0	0	1.000	0.986	0.993	0.986	1.000	1.000	0.000	0.534
(0.400,0.500]	0	0	1.000	0.986	0.993	0.986	1.000	1.000	0.000	0.534
(0.300,0.400]	0	0	1.000	0.986	0.993	0.986	1.000	1.000	0.000	0.534
(0.200,0.300]	0	0	1.000	0.986	0.993	0.986	1.000	1.000	0.000	0.534
(0.100,0.200]	0	0	1.000	0.986	0.993	0.986	1.000	1.000	0.000	0.534
(0.000,0.100]	0	0	1.000	0.986	0.993	0.986	1.000	1.000	0.000	0.534

Fig. 5(c): Evaluation results

5. CONCLUSION

With the rising importance of the web not only as an information portal but also as a business edge, the importance of web usage mining is unquestionable. Web access logs contain abundant raw data that can be mined for web access patterns, which in turn can be applied to improve experience of users. By taking into consideration we have mainly focused on designing of web usage mining system for discovering usage pattern and analyze the pattern to evaluate its precision. Combined techniques of supervised machine learning algorithm is used to discover pattern and analysis removing uncertainty. Experimental results infer increasing accuracy provided by the system and uncertainty is addressed through the usage pattern analysis. User behavior pattern discovery is made using Multiclass decision forest and classification technique with resampling methods of Bagging and Replicate. Accuracy results is found increasing with replicate resampling method.

6. REFERENCES

[1] da Silva EQ, Camilo-Junior CG, Pascoal LML & Rosa TC, "An evolutionary approach for combining results of recommender systems techniques based on collaborative filtering", Expert Systems with Applications, Vol.53, (2016), pp.204-218.

[2] Lopes P & Roy B, "Dynamic Recommendation system using web usage mining for e-commerce users", Procedia Computer Science, Vol.45, (2015), pp.60-69.

[3] Chikhaoui B, Wang S, Xiong T & Pigot H, "Pattern-based causal relationships discovery from event sequences for modeling behavioral user profile in ubiquitous environments", Information Sciences, Vol.285, (2014), pp.204-222.

[4] Belk M, Papatheocharous E, Germanakos P & Samaras G, "Modeling users on the World Wide Web based on cognitive factors, navigation behavior and clustering techniques", Journal of Systems and Software, Vol.86, No.12, (2013), pp.2995-3012.

[5] Liu H & Keselj V, "Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests", Elsevier, (2007), pp.304-330.

[6] Middleton SE, Shadbolt NR & De Roure DC, "Ontological user profiling in recommender systems", ACM Trans. Inf, Syst., Vol.22, (2004), pp.54-88.

[7] Degemmis M, Lops P, Semeraro G & Abbattista F, "Extraction of user profiles by discovering preferences through machine learning", Proceedings of IIS, (2003), pp.69-78.