Dynamic Log Session Identification Using A Novel Incremental Learning Approach For Database Trace Logs

D.Kavitha, B.Kalpana

Abstract—: Identification of session is very significant task in database for determining helpful patterns from database trace logs files. In recent years, several number of work have been done to dynamic log session identification among them n-gram models is recently proposed, which produces higher log session identification results. But the major issue of the n-gram model is that it assumes the entire database query to be static, so dynamic query type is not applicable. In this paper, a novel online incremental learning for dynamic log data trace session identification schema based on the adaptation method to database trace logs is proposed. It is applied for dynamic log session identification with automatic selection of threshold based on the standard deviation schema, so it is named as DS-OILSD. The proposed novel DS-OILSD schema is varied from normal n-gram model since the proposed work consists of two major modifications. The first one is to solve the parameters adjustment problem of the IL in online and offline manner incrementally. The second one is used to dynamic management of query types and allocating initial probabilities to the n-grams models. The proposed DS-OILSD leaning method is based on modified MAP estimation schema for dynamic changing adaptation of the query types and is instinctively reasonable. It directly solves the problem of dynamic log session identification, in which three types of learning are performed such as labeled data, semi-labeled, unlabeled data for various categories of training data. Finally experimental work is conducted to proposed and existing state of art schema for dynamic log session identification and experimentation results are evaluated based on the parameters like, F-measure, and precision for clinic database log files.

Index terms - Database trace logs, online incremental learning, Statistical language modeling, Session identification, standard deviation, n-gram model.

1 Introduction

Due to the growth and development of the internet, the usage of web has increased without any proper management of infrastructure. In addition management of, plenteous information on the web and heterogeneous environment of the web has become a very tedious task, because of several numbers of users. Most important and significant area to mine user behavior is web usage mining [3]. The major objective of the web usage mining is to discover the navigation patterns of the user from web log data. Even though identifying the user interest from web log data is not an easy task and tremendously confusing, since it consists of several number of the navigation patterns, there remains a possible design for discovering valuable information from the interactions among a web site and its users.

In general, web log record [4] of user is stored based on the user requested query from the web server. Log file consists of several numbers of data for each user; it includes the IP address of the system used by user, the web user starting and completion time along with data, image or specific information which are requested by web user, and so on. Based on the usage of the websites, a web logs record log file consists of thousands, hundreds and millions of data for each user on each day. So finding the navigation patterns from this information becomes more difficult, so the log entries of each web user are clustered into different sessions known as web log session. To manage and store the information of the web log user, database system [5] have been focused in recent years, because of the uniqueness of its hardware and software mechanism. The database system is maintained based on the number of request made by the user on the specific period of the time. So the examination of the workload on the database system plays very imperative role. To quickly examine the database, automatic discovery of the web log session identification is performed to find consequential patterns and relationships from voluminous data stored in the database of web server log files.

Most of the recent work in database system is performed based the searching behavior of database users [6-7]. To examine the web log session and discover patterns from database log files become tedious tasks, since each user have maintain multiple sessions for the duration of specific time. To conquer this problem automatic session identification is performed based on the timeout schema, where the session is differentiated based on the time interval with predefined threshold value, but it becomes very difficult to set threshold for each session identification process.

- *D.Kavitha is currently pursuing part time PhD Degree program and working as Assistant Professor in KGiSL Institute of Information Management, India, PH-09842866416. E-mail:kavithadevaraj76@gmail.com*

- *Dr.B.Kalpana is currentlyworking as Professor in Department of Computer Science in Avinashilingam Institute of Home Science, India, PH-09486447430. E-mail:kalpanabsekar@gmail.com*

The examination of Web log [9] may possess several challenges on information search and security. The discovered navigation patterns might be useful to discover the searching behavior based on the number of queries which is requested by user. In recent work prediction has been done based on the prefetching predicted queries, and cache replacement [9] schema is proposed and automatic web log session identification is done based on the n-gram model [10], which uses entropy based threshold schema for test samples. The major issue of the statistical n-gram model is that it is not applicable for dynamic query type.

The objective of this paper is to solve the dynamic log session identification method problems by proposing novel dynamic web log session identification with online incremental learning method for database trace files and applied to web usage mining. The proposed DS-OILSD (Dynamic Session-Online Incremental Learning based on Standard Deviation) is performed based on the statistical probabilistic method. The proposed DS-OILSD might not depend on the time base session identification, instead uses information which is searched by the user for automatic dynamic web session identification ,in addition it measures the standard deviation threshold values to select the test information for session identification. To determine accuracy and effectives of the proposed DS-OILSD method and existing statistical information theoretical n-gram analysis model, the experimentation work is conducted to benchmark dataset with traceable database web log files. Experimentation work measure the efficiency based on the general classification parameters.

The remainder of this paper is summarized in the following manner. In first section 2, discusses and describes the information of the web log session identification method for web log data and then finally specify the issues of the present methods. In Section 3, it introduces the proposed online incremental learning for dynamic web log session identification for web user. In Section 5, it introduces the experimentation work; the evaluation is done for existing and proposed schema which is applied for benchmark dataset. In Section 4, it examines and converse the investigational results. Finally, conclude the entire study in Section 5 and issues of the work is also studied in conclusion.

2 Background Knowledge

Web log session identification schema is easily done based on the preprocessing methods. In the preprocessing methods the conventional session identification methods is fully investigated based on dynamic timeout [11]. In initial stage of web log session identification, session timeout is computed for each web log user based on the searched web page with their degree value. Based on the computed session timeout, user sessions are identified through dynamically adjusting the session timeout. [12] Presents new graph tree based web log session identification for user's session through taking into consideration of IP address, web browser, and timeouts between intra and inter session, backward reference examination exclusive of searching the complete tree. In the proposed graph tree based web log session identification schema entire user session sequences is also created. Experimental work on web user performance is measured between reference length and maximal reference schema. The new graph tree based web log session identification achieves higher precision and less time complexity for web session identification.

Examining the searching behavior from the web users and extracting the valuable information based on the user interest becomes important task for Web usage mining. In order to perform this task the continuously monitoring the web user search behaviors is represented in the form of graph .The most visited websites is represented in the form of graph along with user given keyword so it is named as site-keyword graph [13]. The site-keyword graph identifies the user's interest for each keyword in their websites. The mining of the information beginning Web logs provides right to use data with the intention of have to be controlling proficiently in order to be capable to make use of them for examination. In order to solve this problem the solution is done based on the database management schema which sustain and handle the required information. The detail description of the proposed database schema is explained in [14] for different users along with recorded web log information. Kapusta et.al [15], proposed a web log session identification schema is done based on the time threshold. To decide time threshold value, in the proposed work, we make use of the Length variable which denotes the total time which is spent by user on specific period of the time on a particular site. The experimentation work of the proposed time threshold is compared with existing rule based session identification for web usage mining.

3 Proposed Incremental Learning Approach For Dynamic Web Log Session Identification

This work presents a novel dynamic web log session identification schema based on the language model, and is different from existing methods as it is for both static and dynamic queries. The proposed DS identification schema is based on the online incremental learning method and it is applied for real time clinic application to the trace logs. The major objective of the proposed DS-OILSD schema is to find and automate web log session identification based on the borders of user sessions for database trace log files. In addition, the proposed DS-OILSD schema solves the parameter range problem by proposing a standard deviation for automatic threshold. Experimental results are evaluated based on the parameters like, F-measure, standard deviation threshold value and precision for clinic database trace log files of various user sessions.

In proposed model the entire training data samples is categorized into three types. Learning is introduced for training samples under labeled, training and unlabeled samples. Training data is labeled with the use of learning from labeled data (LLD) method. The unlabeled data samples are learned with the use of unlabeled data (LUD) method by considering both intra and inter session timeout for web session identification.

The semi-supervised learning is performed with the use of learning from semi-labeled data (LSD) by calculation of probabilities for labeled training data samples. These three types of models are applied to real database trace log files. As discussed in the proposed DS-OILSD work, it majorly consists of two steps, one for general incremental learning methods for dynamic session identification, and then second step would be the standard deviation based automatic threshold function for web log session identification. The first step of the work uses the general procedure of the incremental learning algorithm for dynamic session identification based on the text corpus TC for each user session with their words$w$.

3.1General Incremental Learning Model For Dynamaic Session Identification Purpose Language Model

In this proposed DS-OILSD, the probability values for each text corpus Tc with word is represented as $\mathrm{Tc} = \mathrm{w}_1, .. \mathrm{w}_\mathrm{n}$, which is determined as follows ,

$$P(Tc) = P(w_1)P(w_2|w_1)\prod_{i=2} P(w_i|w_{i-1}, w_{i-2}) \qquad (1)$$

The goal of this proposed DS-OILSD is to discover the similar user session identification and identify the dynamic session for each text corpus Tc for input X

$$\widehat{Tc} = \arg\max P(X|Tc)P(Tc) \qquad (2)$$

In this paper proposed DS-OILSD framework uses the concept of Katz smoothing [18] for web log session identification which is defined as follows

$$P_{katz}(w_i|w_{i-2}, w_{i-1}) = \begin{cases} \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})} & if\ r > r_T \\ d_r \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})} & if\ 0 < r \leq r_T \end{cases} \qquad (3)$$

where $C(\cdot)$ is denoted as the total number of the count for specific event of web log session series , r is denotes as the number of times the selected word is occurring the text corpus , $\alpha(w_{i-2}, w_{i-1})$ and d_r be the smoothing parameters for web log session identification . It is dynamically changed if the searching is increases than the training dataset samples for web log session identification ,to solve this problem MAP estimation is done for dynamic changes of session which is written as follows,

$$\phi = \{\lambda_{hw} | w \in W, h \in H\} \qquad (4)$$

where W denotes the set of all possible words from the text corpus of user session trace log files and H is denoted as the set of user database trace log files histories {h} , ϕ is defined as the web log session identification observed results for X data , which maximizes the posterior probability P(ϕ|X), is described as follows,

$$\phi_{MAP} = \arg\max_{\phi} P(X|\phi)P(\phi) \qquad (5)$$

where X is the training dataset samples which is collected from database trace log files which is used for MAP adaptation and P(ϕ)posterior probability for observed samples is described as follows,

$$P(\phi) = k \prod_{w \in W} \prod_{h \in H} \lambda_{hw}^{\alpha_{hw}-1} \qquad (6)$$

where k is a constant value and $\alpha_{hw} - 1$ is denoted as the hyper parameter for dynamic web log session identification from n-gram model (hw) occurring word count for specific user session event TC_{hw}^{T} in the training dataset samples as

$$\alpha_{hw} = C_{hw}^{T} + 1(w \in W, h \in H) \qquad (7)$$

λ_{hw} is determined as ,

$$\lambda_{hw} = \frac{C_{hw}}{\sum_{w \in W} C_{hw}} \qquad (8)$$

Known text corpus data for user session database trace log files X, $C_{hw}^{(A)}$ is the occurring word count value for web log user session history (hw) ,is described as follows ,

$$P(X|\phi) = k \prod_{w \in W} \prod_{h \in H} \lambda_{hw}^{C_{hw}^{(A)}} \qquad (9)$$

From the above equations ,equation (5) is reformulated as ,

$$\phi_{MAP} = \arg\max_{\phi} k \prod_{w \in W} \prod_{h \in H} \lambda_{hw}^{C_{hw}^{(T)} + C_{hw}^{(A)}} \qquad (10)$$

$$\sum_{w \in W} \lambda_{hw} = 1 \qquad (11)$$

$$\lambda_{hw}^{MAP} = \frac{C_{hw}^{(T)} + C_{hw}^{(A)}}{\sum_{w \in W}(C_{hw}^{(T)} + C_{hw}^{(A)})} \qquad (12)$$

During the web search engine, assumes it that the present user modifies some words $w_1, \ldots w_n$ in the user session sequence, but the percentage value for session identification is also highly changed, so in order to manage this problem the original P(ϕ)posterior probability for observed samples is reformulated and modified based on the changed word counts, consequently Equation (12) be supposed to be modified as,

$$P(w|h) = \frac{C_{hw}^{(T)} + \alpha * C_{hw}^{(A)}}{\sum_{w \in W}\left(C_{hw}^{(T)} + \alpha * C_{hw}^{(A)}\right)} \qquad (13)$$

From the equation (13) the parameter α will be changed automatically when the number of the session and queries given by user is increased in dynamic manner .So the parameter α will be determined based on the highest probability value for dynamic session identification , which is determined using the following equation (14),

$$\alpha = \begin{cases} \alpha_{min} if \alpha_{good} \leq \alpha_{min} \\ \alpha_{good} if \alpha_{min} < \alpha_{good} < \alpha_{max} \\ \alpha_{max} if \alpha_{good} \geq \alpha_{max} \end{cases} \quad (14)$$

where α_{min} and α_{max} is denoted as the maximum and minimum value for dynamic web log session identification which is decided based on the automatic threshold function .In this work the automatic threshold function is determined based on the standard deviation ,which is calculated based on the highest posterior probability value determined from MAP estimation function for both online and offline manner.

3.2 Automatic Selection of Threshold for Dynamic Session Identification Based on the Standard Deviation:

To know and find the importance value of the automatic standard deviation threshold value, we perform the experimentation work between the proposed model and it is compared with existing methods and result is shown in Fig.3 and Fig.4. In the experimentation work the standard deviation value of the present session and next session deviation is measured based on the following function defined as,

$$\frac{SD(S_1) - SD(S_0)}{SD(S_0)} \quad (15)$$

where S_0 is denotes as the series of request which is given by user in specific time and S_1 is denotes as the series of request which is given by user and followed from S_0session of test samples. In equation (15), the automatic standard deviation threshold value is determined as follows,

$$\sigma = \sqrt{\frac{\sum\left(P(w|h) - \overline{P(w|h)}\right)^2}{n}} \quad (16)$$

Where σ is the standard deviation for each user session, $P(w|h)$be the probability value for single user session $\overline{P(w|h)}$ mean probability value for all user session for identification, n be the number of session in the database trace log files.

4 Performance Evaluation For Database Trace Logs

In this section, we evaluate novel dynamic log session identification schema through comparing the accuracy of the existing schemas such as dynamic n gram analysis, timeout method. The experimental work on proposed DS-OILSD with n gram model, existing n-gram model and timeout threshold model is conducted on clinical database trace log files. The database trace log files of clinical database are collected through via Microsoft SQL Profiler. It is capable of continuously monitoring all queries which are given and submitted by client application within a specific period of time. The database trace log contains the information of client application with the size of 400M bytes with 81,417 queries submitted by user regarding

total nine types of application .These queries submitted by user regarding applications is specified as front-end sales, report generation in daily, report generation per month, information backup, and system management. In initial stage preprocessing is done to reduce the reputation of the queries for same users. By this step we attain 7,244 SQL queries for front-end application. The preprocessed dataset is categorized into the major dataset and select those dataset for testing purpose which is named as D_1,D_2,D_3 &D_4 D4. The detail description of this data set is specified in Table 1.

TABLE 1
TESTING DATA SETS

Name	Number of queries	Number of sessions	Average session length
$\boldsymbol{D_1}$	677	65	10.4
$\boldsymbol{D_2}$	441	56	7.9
$\boldsymbol{D_3}$	816	118	6.9
$\boldsymbol{D_4}$	1181	153	7.7

In both proposed queries submitted by user regarding model the entire training data samples is categorized into three types namely training , labeled training and unlabeled samples.

In order to measure the performance accuracy of the dynamic web session identification methods, we use a first most performance metric as F-measure. The mean average value of precision and recall results for web session identification of test sequence is named as the F-Measure. The precision is defined as the dynamic session identification based on the percentage of the truly identified web session correctly to the entirety number of estimated web session. The recall is defined as the percentage of truly identified web session correctly, so the F-measure is defined as follows

$$F - measure = \frac{2 \times precision \times recall}{precision + recall}$$

The unlabeled data samples are learned with the use of unlabeled data (LUD) method .So the performance of the LUD with OILSD and LUD with n-gram model method, is evaluated between three test dataset samples, the trained dataset samples is also used for test data for another samples, the results of F-measure for various test dataset samples are shown in table 2.

TABLE 2
THE PERFORMANCE OF THE LUD WITH OILSD AND LUD WITH N-GRAM MODEL METHOD

Test data	Number of sessions			F-Measure		F-Measure		F-Measure	
	True	Estimated	Correct	Auto SD	OILSD n-gram	Auto entropy	n-gram	Average	Best timeout
D_2	58	60	52	0.895	0.91	0.84	0.78	0.78	0.71
D_3	120	108	98	0.914	0.92	0.845	0.80	0.80	0.62
D_4	158	145	112	0.82	0.83	0.69	0.65	0.64	0.69

Table 2 show the F-measure results of various test dataset samples in terms of the Number of sessions which is categorized into three ways such as truly, estimated and corrected. The F-measure values of LUD with OILSD and LUD with n-gram model method is shown in Table 2. Note that the LUD with OILSD achieves higher F-measure value for various datasets.

Training data is labeled with use of learning from labeled data (LLD) method .So the performance of the LLD with OILSD and LLD with n-gram model method, is evaluated between four test dataset samples, the trained dataset samples is also used for test data for another samples, the results of F-measure for various test dataset samples are shown in Table 3.

TABLE3

THE PERFORMANCE OF THE LLD WITH OILSD AND LLD WITH N-GRAM MODEL METHOD

Test data	Number of sessions			F-Measure		F-Measure		F-Measure	
	True	Estimated	Correct	Auto SD	OILSD n-gram	Auto entropy	n-gram	Average	Best timeout
D_1	69	68	62	0.93	0.94	0.89	0.80	0.80	0.70
D_2	59	66	58	0.92	0.923	0.88	0.88	0.81	0.71
D_3	121	108	98	0.934	0.94	0.84	0.85	0.78	0.62
D_4	162	169	138	0.842	0.846	0.78	0.81	0.73	0.69

Table 3 shows the F-measure results of various test dataset samples in terms of number of sessions which is categorized into three ways such as truly, estimated and corrected. The F-measure values of LLD with OILSD and LLD with n-gram model method is shown in Table 3. Note that the LLD of OILSD-n gram model achieves higher F-measure value for various datasets which is higher than the Best timeout and automatic entropy method.

The semi-supervised learning is performed with the use of learning from semi-labeled data (LSD) by calculation of probabilities for labeled training data samples all of test samples. In this method the dataset one is used as the tune data and the remaining dataset samples is used as test dataset samples to calculate probabilities value and shown in Table 4.

TABLE 4

THE PERFORMANCE OF THE LSD WITH OILSD AND LSD WITH N-GRAM MODEL METHOD

Test data	Number of sessions			F-Measure		F-Measure		F-Measure	
	True	Estimated	Correct	Auto SD	OILSD n-gram	Auto entropy	n-gram	Average	Best timeout
D_2	58	55	52	0.91	0.92	0.88	0.80	0.80	0.71
D_3	122	124	112	0.92	0.938	0.88	0.80	0.80	0.62
D_4	155	152	121	0.83	0.821	0.78	0.70	0.70	0.69

Fig. 2, shows the precision results of various LUD, LLD and LSD methods with n gram model for entropy and standard deviation, here both standard deviation and entropy function is used for dynamic automatic web session identification of database trace log files .Fig.2 also measure the precision results for timeout threshold with four different datasets.

Fig. 1, shows the F-measure performance accuracy results of various LUD, LLD and LSD methods in with n gram model for entropy and standard deviation is given. Here both standard deviation and entropy function is used for dynamic automatic web session identification of database trace log files. Fig.1 also measures the F-measure results for timeout threshold with four different datasets.

From Fig.1 observed and experimented that the dynamic web log session identification for standard deviation with three n –gram models of incremental learning produces best results since the proposed OILSD n-gram methods dynamic and static queries are supported simultaneously, when compare to LSD-SD,LLD-SD produces higher F-measure results for unlabeled samples for all dataset ,LUD-SD performs slight worst results when compare to LSD-SD and LLD-SD.

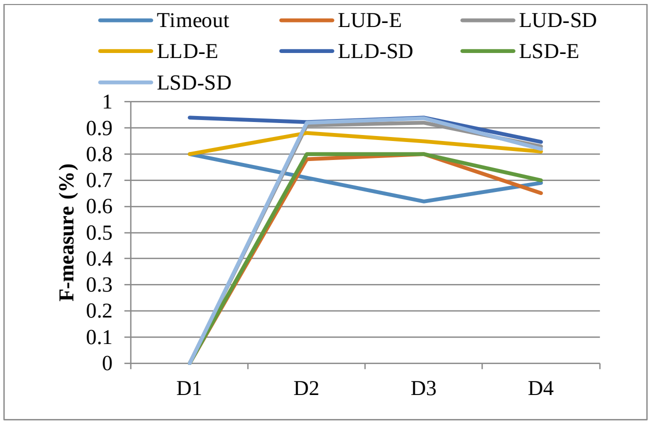

Fig. 1 Comparison of all the methods vs F-measure

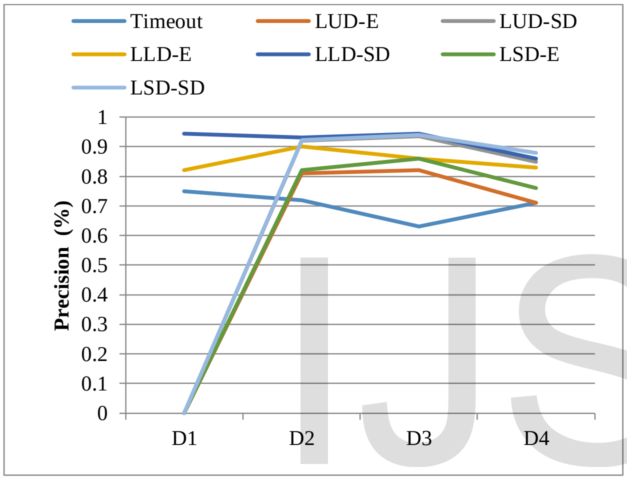

Fig. 2 Comparison of all the methods vs Precision

From Fig.2 observed and experimented that the dynamic web log session identification for standard deviation with three n –gram models of incremental learning produces best precision results, since the proposed OILSD n-gram methods dynamic and static queries are supported simultaneously, when compared to LSD-SD, LLD-SD produces higher precision results for unlabeled samples for all dataset, LUD-SD performs slight worst results when compare to LSD-SD and LLD-SD.

Fig. 3, shows the F-measure performance accuracy of the standard deviation with n gram model and it is largely depend on the threshold function for automatic web session identification of database trace log files.

For example if there are m session with n different queries which is submitted by each user the standard deviation method is more applicable ,then final standard deviation of each value is determined based on the sort the standard deviation in either ascending or descending order. It shows that the threshold is increases for n session then F measure results are slightly decreases.

Fig. 4 shows the precision results of the standard deviation with n gram model and it is largely depend on the threshold function for automatic web session identification of database trace log files.

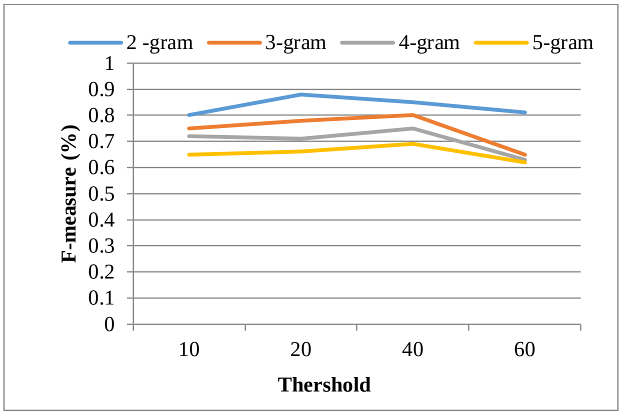

Fig.3 F-measure accuracy for various standard deviation threshold values

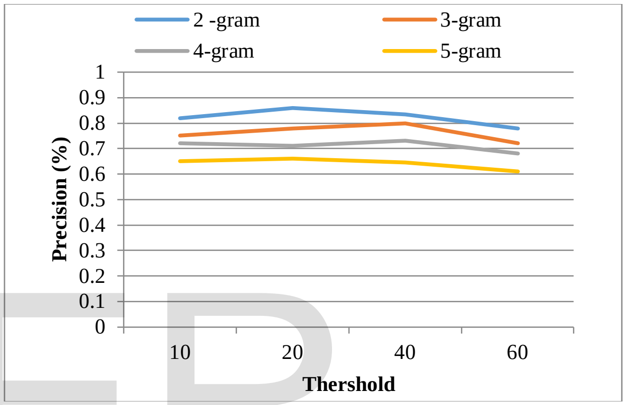

Fig.4 Precision accuracy for various standard deviation threshold values

For example if there are m session with n different queries which is submitted by each user the standard deviation method is more applicable ,then final standard deviation of each value is determined based on the sort the standard deviation in either ascending or descending order. It shows that the threshold is increases for n session then precision result is slightly decreases.

5 Conclusions And Future Work

In this paper we presents a novel dynamic log session identification schema, which is proposed for log data traces files form clinic database. The proposed work presents a novel dynamic log session identification schema based on the online incremental learning algorithm to trace the log files of database system. In the proposed incremental learning schema is different from existing log session identification schema since the proposed DS-OILSD basically assumes that all the databases query patterns becomes dynamic, motivation be further desirable if the dynamic query patterns is changed continuously for web log database trace files. In the proposed DS-OILSD model, the dynamic web session identification is automatically selected based on the standard deviation threshold value and automatic tuning of the web log session for OILSD schema; it is applied for log data traces files from clinic database. In the proposed DS-OILSD schema, MAP estimation is done for dynamic changing adaptation of query types based on the dynamic weight changes in the language model. The MAP estimation schema adjusts the parameters DS-OILSD schema capably, and keeps away from sharp changing to formulate the form constant. The exposed patterns are able to be second-hand to forecast incoming queries relying on the queries previously presented, which is able to use to enhance the database performance through effective query method and cache substitution. The method is established to be considerably better than the existing n gram model designed for dynamic log session identification. The experimentation work is measured based on the performance measures namely, F-measure and precision to evaluate the performance of dynamic log session identification. The extension of the present work will be to analyze the accuracy of the proposed incremental learning and automatic standard deviation threshold log session identification method to other types of log data based benchmark dataset and DNA sequence analysis, and examining the efficiency of the proposed novel incremental learning approach for those datasets using the classification parameters. An automatic threshold of the standard deviation threshold setting for log session identification is extended based on the evolutionary algorithm to select the threshold value.

REFERENCES

[1.] Eirinaki, Magdalini, and Michalis Vazirgiannis. "Web mining for web personalization." *ACM Transactions on Internet Technology (TOIT),* vol.3, no.1 , pp.1-27, 2003.

[2.] Khribi, Mohamed Koutheaïr, Mohamed Jemni, and Olfa Nasraoui, "Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval" *Advanced Learning Technologies, 2008. ICALT'08. Eighth IEEE International Conference on*, pp. 241-245, 2008.

[3.] Liu, Bing, "Web usage mining" Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data pp.449-483, 2007.

[4.] L. J.Grace, V. Maheswari, & D. Nagamalai, Web Log Data Analysis and Mining. *Advanced Computing Communications in Computer and Information Science,* vol.133, pp. 459-469, 2011.

[5.] A., Caplinskas, G.Dzemyda, , & A.Lupeikiene, "Databases and Information Systems" *VII: Selected Papers from the Tenth International Baltic Conference, DB&IS,* vol. 249, 2013.

[6.] JDuyand, LVaughan "Usage data for electronic resources: a comparison between locally collected and vendor-provided statistics", *J Acad Libr*vol.29, no.1, pp.16–22, 2003.

[7.] Q.Yao, A An, "Using user access patterns for semantic query caching", *In: Proceedings of the 14th international conference on database and expert systems applications (DEXA'03), Prague, Czech Republic*, pp.737–746, 2003.

[8.] B. J. Jansen, *Handbook of research on web log analysis*, IGI Global, 2008

[9.] Q.Yao, A, "AnCharacterizing database user's access patterns", *In: Proceedings of the 15th international conference on database and expert systems applications (DEXA'04), Spain,* pp.528–538, 2004.

[10.] Xiangji Huang ,Qingsong Yao, Aijun An , "Applying language modeling to session identification from database trace logs", *Knowledge and Information Systems* ,vol.10, no.4, pp. 473–504, 2006.

[11.] Xinhua, H., & Qiong, W., "Dynamic timeout-based a session identification algorithm", *International Conference in Electric Information and Control Engineering (ICEICE),*pp. 346-349, 2011.

[12.] Arumugam, G., & Suguna, S., "Optimal algorithms for generation of user session sequences using server side web user logs", *In Network and Service Security, 2009. N2S'09. International Conference on* , pp. 1-6, 2009.

[13.] Murata, T., & Saito, K. "Extracting users' interests from web log data", *In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence* , pp. 343-346, 2006.

[14.] Agosti, Maristella, and Giorgio Maria Di Nunzio. "Web Log Mining: A study of user sessions." *Proceedings of the 10th DELOS Thematic Workshop on Personalized Access, Profile Management, and Context Awareness in Digital Libraries (PersDL).* 2007.

[15.] Kapusta, J., Munk, M., Svec, P., & Pilkova, A. (2014). Determining the Time Window Threshold to Identify User Sessions of Stakeholders of a Commercial Bank Portal. *Procedia Computer Science,* vol.9, pp.1779-1790.

[16.] Kapusta, Jozef, Michal Munk, and Martin Drlík. "User Session Identification Using Reference Length." *DIVAI 2012*, 2012.

[17.] Munk, Michal, and Martin Drlik. "Influence of different session timeouts thresholds on results of sequence rule analysis in educational data mining."*Digital Information and Communication Technology and Its Applications. Springer Berlin Heidelberg,*pp. 60-74, 2011

[18.] S.M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Trans. on Acoustics, Speech, andSignal Processing*, vol. 35, no.3, pp.400~401, 1987.