Research Journal of Applied Sciences, Engineering and Technology 8(12): 1472-1479, 2014
DOI:10.19026/rjaset.8.1124
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: August 01, 2014 Accepted: September 22, 2014 Published: September 25, 2014

Research Article

An Improved Web Log Mining and Online Navigational Pattern Prediction

[1]D. Anandhi and [2]M.S. Irfan Ahmed
[1]Department of Computer Technology and Applications, Coimbatore Institute of Technology,
[2]Department of Computer Applications, Sri Krishna College of Engineering and Technology,
Coimbatore-641008, India

Abstract: The aim of this study is to improve web log mining and online navigation pattern prediction. Web mining is an active and wide area which incorporates several usages for the web site design, providing personalization server and other business making decisions etc. Efficient web log mining results and online navigational pattern prediction is a tough process due to vast development in web. It includes the process such as data cleaning, session identification and clustering of web logs generally. In this study initially the web log data is preprocessed and sessions are identified using refined time-out based heuristic for session identification. Then for pattern discovery a density based clustering algorithm is used. Finally for online navigation pattern prediction a new technique of SVM classification is used, which rectifies time complexity with increased prediction accuracy.

Keywords: DBSCAN, optics, support vector machine, web mining

INTRODUCTION

World Wide Web is a huge repository of web pages and links. It provides abundance of information for the Internet users. The development of web is fabulous as approximately one million pages are included daily. Users' accesses are recorded in web logs. Because of vast usage of web, the web log files are growing at a faster rate and the size is becoming huge. Web data mining is the application of data mining techniques in web data. It automatically discovers and extracts information from Web documents and services (Etzioni, 1996). Web Usage Mining applies mining techniques in log data to extract the behavior of users which is used in various applications like personalized services, adaptive web sites, customer profiling, prefetching, creating attractive web sites etc. It consists of main three categories, Web usage mining, Web structure mining and Web content mining. Web usage mining refers to the discovery of user access patterns from Web usage logs. Web structure mining tries to discover useful knowledge from the structure of hyperlinks. Web content mining aims to extract/mine useful information or knowledge from Web page contents.

Web usage mining has many applications (Facca and Lanzi, 2003), e.g., personalization of web content, pre-fetching and caching and support to the design, recommendation systems (Huang *et al*., 2006), prefetching and cashing (Huang and Hsu, 2008), among others. The web usage mining process can be divided into three steps based on Cooley *et al*. (1999). It starts first with data cleaning and pre-processing. Second, the pre-processed data is mined for some hidden and useful knowledge. Finally, the web log mining process ends by analyzing the mining results.

While collecting the data from the web access log, data miners face the following issues to discovery navigation pattern. Differentiating the different visitors because the problem is that certain visitors may use proxy servers or share the same machine to browse the website. Therefore, using the IP address assigned to a user's computer as a unique identifier might lead to erroneous results. Second, when the users use backward and forward buttons of the browser and these actions cannot be recorded in the log. But the missing information is also needed. Also, when a user requests a resource, the server will most likely log more than one entry. Many records might be added to the log for one single request. We have to get rid of the extra information the log collects. Third, we need to identify the different browsing sessions a user might have within a period of time. Forth, we need to estimate the time spent by a user on the last page he/she visited during a specific session. In addition to the aforementioned problems which are directly related to the web usage mining process itself, there are problems related of its applications such as online navigational pattern prediction. This prediction task has to be done in a timely manner with the best accuracy possible (Guerbas *et al*., 2013).

In this study, mainly three processes are focused, namely:

Corresponding Author: D. Anandhi, Department of Computer Technology and Applications, Coimbatore Institute of Technology, Coimbatore-641008, India

- Data cleaning
- Mining navigation pattern
- Predicting online navigation patterns

It starts by creating page views and ends up by generating sessions. In literature, existing approaches for sessions' identification are the time-based heuristics are done. The idea of these time based heuristics is the use of a duration threshold to decide whether a session has ended or not. This work improves such heuristic in order to get better quality results.

DBSCAN is a classic density-based clustering algorithm used for mining navigation patterns. It groups data points which are sufficiently dense into clusters and the discovery process is based on the fact that a cluster can be expanded by any of its core objects. In DBSCAN, the density associated with a point is obtained by counting the number of points in a region of specified radius called Eps around this point. Points with a density above a specified threshold called MinPts are identified as core points. The discovery process starts from an arbitrary point, if it's a core point, the neighborhood query (Viswanath and Babu, 2009) recursively continues and stops at the border points, then another arbitrary ungrouped object is selected and this process is repeated until all data points in the dataset have been placed in clusters or labeled as noise (Guerbas *et al*., 2013).

The classification technique choosing is an important process, here binary classifier SVM is used. It can be efficiently used for linear and non-linear classification of high dimension classification. SVMs are helpful in text and hypertext categorization as their application can significantly reduce the need for labeled training instances in both the standard inductive and transductive settings.

LITERATURE REVIEW

In order to mine for navigational patterns it is mandatory to know what visitors have looked at each time they have visited the website. Each time a visitor comes to the website is considered a session. Identifying users' sessions from the web log is not easy as it may seams. Logs may span long period of time during which visitors may come to the website more than once. Therefore, sessions' identification becomes the task of dividing the sequence of all page requests made by the same user during that period into subsequences called sessions. Many approaches have been used by researchers for sessions' identification. According to (Pabarskaite and Raudys, 2007), the most popular session identification techniques use a time gap between requests.

In Yavas *et al*. (2005), mine the movement patterns of an individual user to form association rules and use these rules to make location prediction. Additionally, they consider the support and confidence in selecting the association rules for making predictions. In Zheng *et al.* (2009a) propose a novel pattern, called Individual Life Pattern, which is mined form individual trajectory data and they uses such pattern to describe and model the mobile users' periodic behaviors. In Monreale *et al.* (2009) proposes a method aiming to predict with a certain level of accuracy the next location of a moving object. The movement patterns extracted for prediction covers three different movement behaviors, including order of locations, travel time and frequency of user visits. In Zheng *et al.* (2009b) Uses a HITS-based model to mine users' interesting location and detect users' travel sequence to make locations prediction and in (Zheng *et al*., 2009b), they consider the location correlation for generating the users' interesting locations and travel sequence.

Spiliopoulou *et al*. (1999) and Spiliopoulou (1999, 2000) focused on the applications of the usage mining. His works on the navigation pattern discovery and web site personalization has special meaning for the e-commerce society and the Web marketplace allocation and will be very helpful for both Web user and administrator. The Web Utilization Miner system is an innovative sequential mining system. Borges and Levene (1999) have explored some algorithms to mine the user navigation pattern and his other papers. He proposed a data mining model to achieve an efficient mining, which captures the user navigation behavior pattern by using Ngrammar approach.

Session identification by referrer has been described in Berendt *et al*. (2003) as follows. Let S be a session under construction and let p and q be two consecutive page requests, where p belongs to S. q is considered to belong to S if its referrer belongs to S and the difference between the timestamps of p and q does not exceed a predefined threshold. Otherwise, q is considered to be part of a new session. According to the comparative study conducted by Berendt *et al*. (2003), session identification by referrer exhibited very poor performance in the presence of framesets. However, time-oriented heuristics were less affected by the presence of framesets.

SVM is a binary classifier has the set of related supervised learning methods used for classification and regression (Vapnik, 1995). They belong to a family of generalized linear classification. A special property of SVM is it simultaneously minimizes the empirical classification error and maximizes the geometric margin. Thus it is called as Maximum Margin Classifiers which follows Structural risk Minimization. SVM map input vector to a higher dimensional space where a maximal separating hyper plane is constructed.

PROPOSED WORK

The WEBMINER is a system that implements parts of this general architecture. The architecture classifies

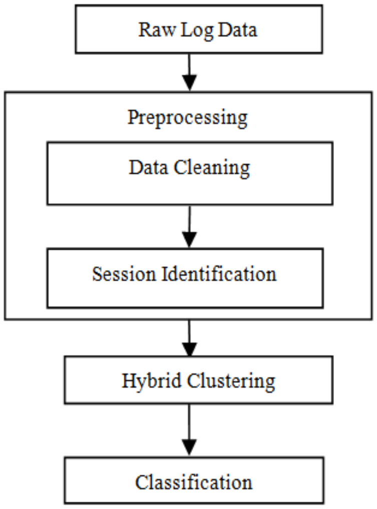

Fig. 1: Proposed work

the Web usage mining process into two main parts. The first part includes the domain dependent processes of transforming the Web data into suitable transaction form. This contains pre-processing, transaction identification and data integration components. The second part contains the largely domain independent application of generic data mining and pattern matching techniques (such as the discovery of association rule and sequential patterns) as part of the system's data mining engine.

The proposed web usage mining approach is shown in Fig. 1. The raw data log which is collect is initially given for pre-processing of web log data, where the unformatted log data is converted into a form that can be directly applied to mining process.

Preprocessing: The pre-processing steps include data cleaning, user identification and session identification. This step consists mainly of removing implicit requests and removing requests made by robots (i.e., ‘‘web crawler’’).

Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data, it detect and remove all major errors and inconsistencies both in individual data sources and when integrating multiple sources. Here it removes implicit requests and robotic requests from the raw data.

User session here is the set of consecutive pages visited by a single user within the duration of one particular visit to a website. For logs that span long periods of time, it is very likely that users will visit the Web site more than once. Once a user has been identified, the click stream of each user is portioned into logical clusters. The method of portioning into sessions is called as Sessionization or Session Reconstruction. There are three methods in session reconstruction. Two methods depend on time and one on navigation.

In this study, we consider a visitor as a robot if one of the two following criteria applies:

- The visitor has requested the file robots.txt
- The visitor exists in the robots lookup list

And in detection of sessions, many techniques are proposed in literature, but time out based technique is widely used by commercial. In this technique a session is identified by the set of requested pages during a predefined threshold time-out interval and to avoid duplications. Reason for opting time-out based sessionizing approach is the fact that this technique can handle a log data on their usual web log format, extended format and also if it used cookies.

The time out interval is a threshold value which is manually given. Based upon the time out the sessions are identified. Shared patterns are identified from the pages used in a session. For example, If in session 1 there is 6 pages say: P1, P4, P14, P20, P421 then in session 2 if there is 3 pages say: P3, P14 and P20. Then the shared patterns are P14 and P20.

Clustering: The DBSCAN technique is a clustering technique used in refining pattern in literature (Reddy and Ussenaiah, 2012). It has wide properties of detects outliers and information regarding number of clusters is not needed. But, it requires another input parameter and it is very tough for the user to guess a good value for that parameters and it is very sensitive to that input parameter. For that reason, use of another algorithm called Ordering Points to Identify the Clustering Structure (OPTICS) which helps in selecting an appropriate value to use as input for DBSCAN. This OPTICS is combined with DBSCAN here to produce optimized clustering results.

OPTICS is an algorithm for finding density-based clusters in spatial data. This OPTICS is used before DBSCAN process. It selects appropriate value to use as input for DBSCAN. The key idea of density-based clustering is that for each object of a cluster the neighborhood of a given radius (ε) has to contain at least a minimum number of objects (MinPts). Thus this technique is hybridize with DBSCAN to propose optimized technique.

The DBSCAN suffers in finding the radius. In some cases, it will be very obvious when clustering users on a map. It is based on a very clever idea: instead of fixing MinPts and the Radius, we only fixMinPts and plot the radius at which an object would be considered dense by DBSCAN. In order to sort the objects on this plot, we process them in a priority heap, so that nearby objects are nearby in the plot.

In this information retrieval system, namely TF-IDF combined with the cosine similarity measure to find the closest sessions to a current online session. To speed up the search process, an inverted index from all sessions is build. We view online pattern detection as the problem of finding the most relevant documents to a query made of set of keywords from a repository of text documents based on the following observations. A text document is made of a collection of some of the words available in the vocabulary of a specific language. Similarly, a session is made of a collection of the references of some of the pages of a website. Text documents may contain repeated terms. Also, a session may contain one page or more many times. This can happen simply when a user reload a page or comes back to one of the pages viewed earlier. Therefore, based on the aforementioned observations, we decide to treat sessions as text documents and benefit from the well developed techniques in the informational retrieval domain to increase results accuracy.

When a visitor is browsing the website, we keep track of the last few requested pages only. We use the concept of sliding window. Each time the visitor requests a new page we slide the window by one. Consequently, the newly requested page will be added and the oldest page in the window will be dropped. Let's assume that the size of the sliding window is w. Before making any prediction about the navigation pattern of the current visitor, wait till the visitor requests at least w pages. Once it is the case, a query vector is created. The size of the vector is the number of pages in the website. The values in the query vector are the TF-IDF values of each requested page existing in the current sliding window. If a page does not exist in the current sliding window then its TF-IDF value is zero. The next step is computing the cosine between the query vector representing the online session of the current visitor and a selected sub-list of vectors representing relevant sessions to the current online session. This can be obtained by the relevant sub-list of vectors by using an inverted index. Therefore, instead of computing the cosine between the current online session and all sessions we do it only for a subset of the sessions.

A TF-IDF value of a specific term in a document d is calculated as the following:

$$\mathrm{TF}(\mathrm{d},\mathrm{t}) = \log\left(1 + \frac{\mathrm{n}(\mathrm{d},\mathrm{t})}{\mathrm{n}(\mathrm{d})}\right) \tag{1}$$

$$\mathrm{TF.IDF}(\mathrm{d},\mathrm{t}) = \frac{\mathrm{TF}(\mathrm{d},\mathrm{t})}{\mathrm{n},\mathrm{t}} \tag{2}$$

where,

n (d, t) = The number of occurrences of term t in document d

n (d) = The number of terms in document d

n (t) = The number of documents containing term t

The cosine between two vectors A, that represents the online query vector and B that represents a vector from the chosen subset of session vectors is calculated as:

$$\cos(\mathrm{A},\mathrm{B}) = \frac{\mathrm{A.B}}{\|\mathrm{A}\|.\|\mathrm{B}\|} \tag{3}$$

The elements of each vector are the TF-IDF values mentioned before.

Support vector machine: SVM is an example of supervised learning classification. It is a binary linear classifier which takes an input and decides to which of the two classes it belongs. The classifier is first trained using a set of training samples. The training samples are pre marked as belonging to one of the two categories and based on these examples, the SVM classifier builds a model that assigns new examples (data to be tested) their suitable classes. The training samples are represented as points in space and are mapped such that there is a clear gap which divides the samples belonging to separate classes. The new samples are then mapped into the same space by analyzing to which of the two classes they suit better (Xing and Guo, 2004).

SVM uses linear model to implement nonlinear class boundaries through some nonlinear mapping the input vectors x into the high-dimensional feature space. The optimal separating hyperplane is determined by giving the largest margin of separation between different classes. For the two-class case, this optimal hyper plane bisects the shortest line between the convex hulls of the two classes. The data are separated by a hyper plane defined by a number of support vectors.

The clustered results under given threshold timeout and above that timeout are given for the classifier.

Initially set of training samples are given, $(\mathrm{x_i}, \mathrm{y_i})$ where the $\mathrm{x_i}$ is the clustered page weights and the $\mathrm{y_i}$ are the signifying ranks that pages belongs to. The two class pattern recognition problem, $\mathrm{y_i} = +1$ or$\mathrm{y_i} = -1$. A training samples $(\mathrm{x_i}, \mathrm{y_i})$ is called clustered under time out if $\mathrm{y_i} = +1$, otherwise it is un clustered pages. SVM constructs a hyper plane that separates two classes and tries to attain maximum separation between the classes. Separating the classes by a large margin minimizes a bound on the projected simplification fault. The simplest representation of SVM is a maximal margin classifier, constructs a linear separator or optimal hyper plane is given by $\mathrm{w\,T\,x} - \gamma = 0$ involving two classes of samples. The gratis parameters are a vector of weights w, which it is orthogonal to the hyper plane.

The SVM attempts to place a linear boundary between the two different classes and orient it in such a way that the margin is maximized. The boundary can be expressed as follows:

$$(\mathrm{w.x}) + \mathrm{b} = 0\,, \mathrm{w} \epsilon\, \mathrm{R^n}, \mathrm{b} \,\epsilon\, \mathrm{R} \tag{4}$$

where, the vector w defines the boundary, x is the input vector of dimension N and b is a scalar threshold.

The optimal hyper plane is required to satisfy the following constrained minimization as:

$$\min\left\{\frac{1}{2}\|\mathrm{w}\|^2\right\} \text{ with } \mathrm{y_i}(\mathrm{w}\,.\,\mathrm{x_i} + \mathrm{b}) \geq 1, \mathrm{i} = 1,2,\ldots\ldots,\mathrm{l}, \tag{5}$$

where, l is the number of training sets.

Using the trained structure the testing is done. The testing produces the binary result.

Here, thus sample of clustered pages under the time out and un-clustered pages above the timeout are given for training. The SVM classifier is trained based on it and test the results for remaining pages. This makes the system to classify the pages and order them easily which reduces the time.

RESULTS AND DISCUSSION

This section gives brief description about the dataset used for this technique and the statistical analysis of the results obtained by this technique which is compared with some existing works. And shown the gain earned in terms of running time using our proposed SVM classification to predict online navigational patterns.

The real log files taken from various computers are taken to evaluate here. The proposed and existing work here is done in ASP.NET platform in Windows XP OS.

The proposed work is done in three main phases of data cleaning, user identification and clustering. The performance of this work is evaluated for three different time outs.

The Table 1 provides the sample timeout for the sessions shared obtained from the existing and proposed techniques. And improvement (%) obtained by the proposed technique while compared with the existing approach is also given. The obtained improvement for timeout of 10 and 20 gives 0.61% where else timeout 30 gives 2.03% of improvement.

Ordered weights of pages present for different timeout of 10, 20 and 30 min, respectively for existing KNN based pattern mining and proposed SVM based pattern mining is shown below.

The above Table 2, gives the page weight for the timeout 10 before SVM process of the proposed process which is compared with existing technique. The weight of ten pages is listed with its obtained page rank by the existing and proposed technique. The frm page takes weight of 0.342, admin's page weight is 0.249, loginentry page is 0.172, user_login page is 0.151, webresource page is 0.135, admin_home is 0.364, guestlogin is 0.156 and default is 0.142.

The above Table 3, gives the page weight for the timeout 20 before SVM process of the proposed process which is compared with existing technique. The weight of eight pages is listed with its obtained page rank by the existing and proposed technique. The frm page takes weight of 0.342, admin's page weight is 0.172, loginentry page is 0.169, admin_home page is 0.364, projectlist page is 0.249, guestlogin is 0.156, user_login is 0.151 and default is 0.142.

Table 1: Sessionizer result of the existing technique for different timeout

	Existing technique			Proposed technique				
Timeout	Correctly identified sessions	False positives	Ratio of false positives/total (%)	Correctly identified sessions	False positives	Ratio of false positives/total (%)	Total identified sessions	Improvement (%) of proposed
10	3675	186	4.81	186	164	4.20	3861	0.61
20	3564	148	3.98	148	125	3.37	3712	0.61
30	2924	187	6.02	187	124	3.99	3111	2.03

Table 2: Comparison of page weight for the pages using timeout 10

Pages	Page weight for timeout 10	Existing technique page rank	Proposed technique page rank
Frm_project_entry.aspx	0.342	1	1
Projectlistedit.aspx	0.249	2	3
Admin_login.aspx	0.172	3	4
Loginentry.aspx	0.169	4	2
User_login.aspx	0.151	5	5
WebResource.axd	0.135	6	6
Admin_home.aspx	0.364	7	8
Guestlogin.aspx	0.156	8	9
Default.aspx	0.142	9	7

Table 3: Comparison of page weight for the pages using timeout 20

Pages	Page weight for timeout 20	Existing technique page rank	Proposed technique page rank
Frm_project_entry.aspx	0.342	1	1
Admin_login.aspx	0.172	2	3
Loginentry.aspx	0.169	3	2
Admin_home.aspx	0.364	4	5
Projectlistedit.aspx	0.249	5	6
Guestlogin.aspx	0.156	6	7
User_login.aspx	0.151	7	8
Default.aspx	0.142	8	4

Table 4: Comparison of page weight for the pages using timeout 30

Pages	Page weight for timeout 30	Existing technique page rank	Proposed technique page rank
Frm_project_entry.aspx	0.342	1	1
Admin_login.aspx	0.172	2	2
Loginentry.aspx	0.169	3	3
Admin_home.aspx	0.364	4	5
Projectlistedit.aspx	0.249	5	4

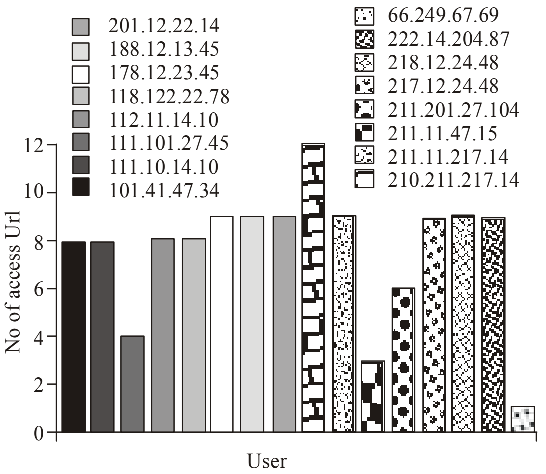

Fig. 2: Report of unique URL obtained after session identification

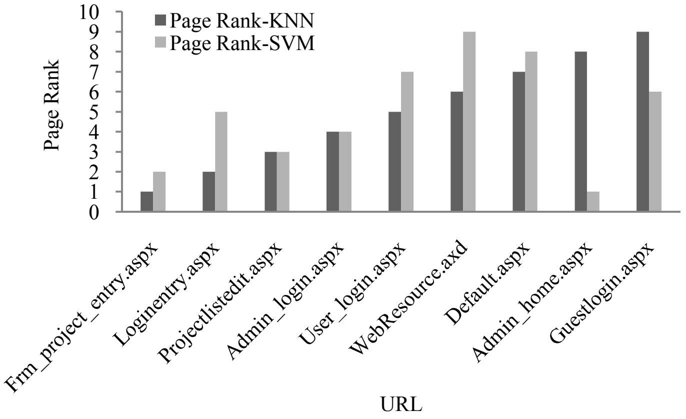

Fig. 3: Comparative page rank for the pages under timeout 10 min

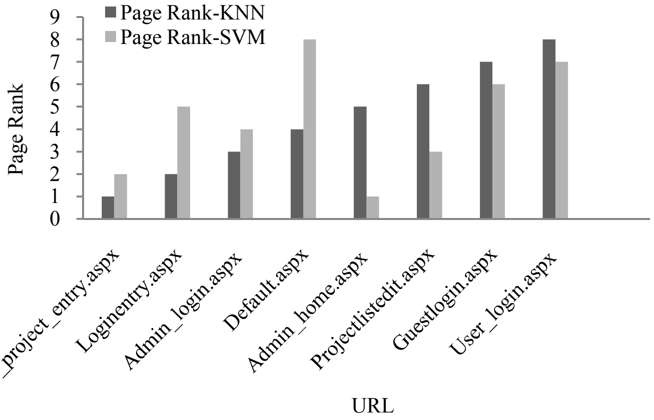

Fig. 4: Comparative page rank for the pages under timeout 20 min

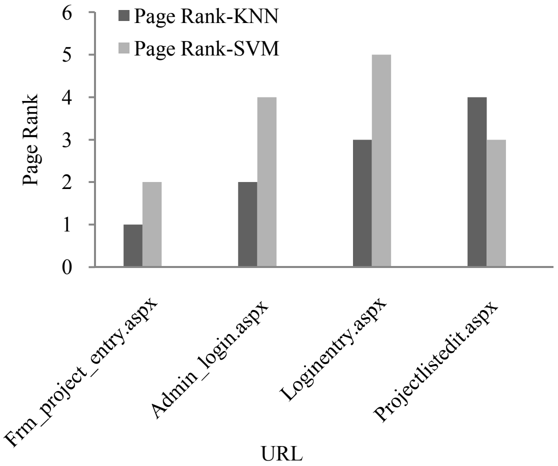

Fig. 5: Comparative page rank for the pages under timeout 30 min

The above Table 4, provides the page weight for the timeout 30 before SVM process of the proposed process which is compared with existing technique. The weight of five pages is listed with its obtained page rank by the existing and proposed technique. The frm page takes weight of 0.342, admin's page weight is 0.172, loginentry page is 0.169, admin_home is 0.364 and project list edit is 0.249.

The above Fig. 2 shows the unique URL obtained for various users after the completion of session identification process.

The above Fig. 3 illustrates the comparative page rank given by existing KNN and proposed SVM for the pages under time out of 10 min.

The above Fig. 4 shows the comparative page rank given by existing KNN and proposed SVM for the pages under time out of 20 min.

The above Fig. 5 displays the comparative page rank given by existing KNN and proposed SVM for the pages under time out of 30 min.

CONCLUSION

Web log mining is an active work which is need of efficient mining of web logs and for online navigational behavior prediction. In this study, the web logs are initially preprocessed by cleaning the data and preparing for clustering. Then from it sessions are identified by using DBSCAN with OPTICS. In this study time out based technique is modified to perform additionally to detect time spend on each page by the visitors. SVM classification is used as a proposed technique here which reduces the time requirement for the clustering and shows the fast convergence of clusters and also improves the quality of clustering for user navigation pattern in web usage mining systems. Experimental results will prove its efficiency of predicting the navigation patterns compared with existing work for the input dataset.

In future more new parameters and features can be used to increase the efficiency of the proposed algorithm. Moreover, work can be extended to modify the TF-IDF values to include the time spent on a page and not only how many times the page has been requested in a session.

REFERENCES

Berendt, B., B. Mobasher, M. Nakagawa and M. Spiliopoulou, 2003. The impact of site structure and user environment on session reconstruction in web usage analysis. In: Masand, B., M. Spiliopoulou, J. Srivastava and O.R. Zaiane (Eds.), WEBKDD 2002 Web Mining for Usage Patterns and User Profiles. LNAI 2703, Springer-Verlag, Berlin, Heidelberg, pp: 159-179.

Borges, J. and M. Levene, 1999. Data mining of user navigation patterns. Proceeding of Revised Papers from the International Workshop on Web Usage Analysis and User Profiling (WEBKDD '99), pp: 31-39.

Cooley, R., B. Mobasher and J. Srivastava, 1999. Data preparation for mining World Wide Web browsing patterns. Knowl. Inf. Syst., 1(1): 5-32.

Etzioni, O., 1996. The World Wide Web: Quagmire or gold mine. Commun. ACM, 39(11): 65-68.

Facca, F.M. and P.L. Lanzi, 2003. Recent developments in Web Usage Mining research. In: Kambayashi, Y., M. Mohania and W. Wob (Eds.), DaWaK 2003. LNCS 2737, Springer-Verlag, Berlin, Heidelberg, pp: 140-150.

Guerbas, A., O. Addam, O. Zaarour, M. Nagi, A. Elhajj and M. Ridley, 2013. Effective web log mining and online navigational pattern prediction. Knowl-Based Syst., 49: 50-62.

Huang, Y.F. and J.M. Hsu, 2008. Mining web logs to improve hit ratios of prefetching and caching. Knowl-Based Syst., 21(1): 62-69.

Huang, Y.M., Y.H. Kuo, J.N. Chen and Y.L. Jeng, 2006. NP-miner: A real-time recommendation algorithm by using web usage mining. Knowl-Based Syst., 19(4): 272-286.

Monreale, A., F. Pinelli, R. Trasarti and F. Giannotti, 2009. WhereNext: A location predictor on trajectory pattern mining. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD' 2009), pp: 637-646.

Pabarskaite, Z. and A. Raudys, 2007. A process of knowledge discovery from web log data: Systematization and critical review. J. Intell. Inf. Syst., 28: 79-104.

Reddy, B.G.O. and M. Ussenaiah, 2012. Literature survey on clustering techniques. IOSR J. Comput. Eng., 3(1): 01-12.

Spiliopoulou, M., 1999. Data mining for the Web. Proceeding of 3rd European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'99), pp: 588-589.

Spiliopoulou, M., 2000. Web Usage Mining for Web site evaluation. Commun. ACM, 43(8): 127-134.

Spiliopoulou, M., L.C. Faulstich and K. Winkler, 1999. A data miner analyzing the navigational behaviour of web users. Proceeding of the Workshop on Machine Learning in User Modelling of the ACAI99.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York, Inc., New York, ISBN: 0-387-94559-8.

Viswanath, P. and V.S. Babu, 2009. *Rough*-DBSCAN: A fast hybrid density based clustering method for large data sets. Pattern Recogn. Lett., 30(16): 1477-1488.

Xing, F. and P. Guo, 2004. Classification of stellar spectral data using SVM. Proceeding of International Symposium on Neural Networks (ISNN'2004). LNCS 3173, Springer-Verlag, Berlin, Heidelberger, pp: 616-621.

Yavas, G., D. Katsaros, Ö. Ulusoy and Y. Manolopoulos, 2005. A data mining approach for location prediction in mobile environments. Data Knowl. Eng., 54(2):121-146.

Zheng, Y., L. Zhang, X. Xie and W.Y. Ma, 2009a. Mining correlation between locations using human location history. Proceeding of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '09), pp: 472-475.

Zheng, Y., L. Zhang, X. Xie and W.Y. Ma, 2009b. Mining interesting locations and travel sequences from GPS trajectories. Proceeding of the 18th International Conference on World Wide Web (WWW'2009), pp: 791-800.