Spatio-temporal Similarity of Web User Session Trajectories and Applications in Dark Web Research

Sajimon Abraham[1] and P. Sojan Lal[2]

[1] School of Management & Business Studies, Mahatma Gandhi University, Kerala, India
[2] School of Computer Sciences, Mahatma Gandhi University, Kerala, India
sajimabraham@rediffmail.com, padikkakudy@gmail.com

Abstract. Trajectory similarity of moving objects resembles path similarity of user click-streams in web usage mining. By analyzing the URL path of each user, we are able to determine paths that are very similar and therefore effective caching strategies can be applied. In recent years, World Wide Web has been increasingly used by terrorists to spread their ideologies and web mining techniques have been used in cyber crime and terrorism research. Analysis of space and time of click stream data to establish web session similarity from historical web access log of dark web will give insights into access pattern of terrorism sites. This paper deals with the variations in applying spatio-temporal similarity measure of moving objects proposed by the authors in PAISI 2010, to web user session trajectories treating spatial similarity as a combination of structural and sequence similarity of web pages. A similarity set formation tool is proposed for web user session trajectories which has applications in mining click stream data for security related matters in dark web environment. The validity of the findings is illustrated by experimental evaluation using a web access log publically available.

Keywords: Spatio-temporal Similarity, Web based Intelligence Monitoring, Web Usage Mining, Web User session Trajectory, Dark Web.

1 Introduction

Modern monitoring systems such as GPS positioning and mobile phone networks have made available massive repositories of spatio-temporal data by recording human mobile activities, call for suitable analytical methods, capable of enabling the development of innovative, location-aware applications. Moving objects which carries location-aware devices, produce trajectory data in the form (*Oid, t, x, y*)-tuples, that contain object identifier and time-space information. In moving object applications, tracking of objects in identifying objects that moved in a similar way or followed a similar movement pattern is to find their common spatio-temporal properties.

With the rapidly increasing popularity of WWW, websites are playing a crucial role to convey knowledge to the end users. In recent years, World Wide Web has been increasingly used by terrorists to spread their ideologies[5] and Web mining techniques have been used in cyber crime and terrorism research [11][12]. Terrorist websites are so dynamic in nature as usually it suddenly emerge, the content and

M. Chau et al. (Eds.): PAISI 2011, LNCS 6749, pp. 1–14, 2011.

hyperlinks are frequently modified, and they may also swiftly disappear [3]. The click stream generated by various users often follows distinct patterns and mining of the access pattern will provide knowledge. This knowledge will help in recommender system of finding browsing pattern of users, finding group of visitors with similar interest, providing customized content in site manager, categorizing visitors of dark web etc.

This paper has made an attempt to establish the resemblance between spatio-temporal similarity of network constrained moving object trajectories[7] and web user session trajectories, which will have wider applications in web based intelligence monitoring and analysis. This approach is unique in the literature illustrating the applications of network constrained moving object trajectories in web usage mining. The proposed spatial similarity includes method for measuring similarities between web pages that has taken into account not only the common URL's visited but also the structural as well as sequence alignment. Based on this a spatio-temporal similarity measure algorithm (WEBTRASIM) is proposed for similarity set of web user session trajectories. We are going to discuss two major topics in detail (i) Trajectory Similarity of Network Constrained Moving Objects which was discussed in PAISI 2010 by the authors of this article[7] and (ii) Similarity of web user session trajectories. Both concepts have wider applications in the domain of Security Informatics. The first one has applications in studying the massive flow of traffic data to monitor the traffic flow and discover traffic related patterns where as the second has applications in tracking terrorist web site visits in the emerging security informatics area of Dark Web Research[12].

The rest of the paper is organized as follows. Related work and literature are surveyed in section 2. We propose a spatio-temporal trajectory similarity measure for web user session trajectory and introducing the algorithm WEBTRASIM in section 3. The implementation and experimental evaluation of this algorithm is made in section 4 and section 5 concludes the paper with directions for future research.

2 Related Work

Most of the works on trajectory similarity of moving objects are inappropriate for similarity calculation on road networks since they use the Euclidian distance as a basis rather than the real distance on the road network [6]. This point has motivated to propose a similarity measure based on the spatio-temporal distance between two trajectories [4] using the network distance. They have proposed an algorithm of similar trajectory search which consists of two steps: a filtering phase based on the spatial similarity on the road network, and a refinement phase for discovering similar trajectories based on temporal distance. In the above work the authors propose a similarity measure based on Points Of Interest (POI) and Time Of Interest(TOI) to retrieve similar trajectories on road network spaces and not in Euclidean space. One of the recent works in trajectory similarity problem for network constrained moving objects [9] introduces new similarity measures on two trajectories that do not necessarily share any common sub-path. They define new similarity measures based on spatial and temporal characteristics of trajectories, such that the notion of similarity in space and time is well expressed, and more over they satisfy the metric properties. The above

work has identified some of the drawbacks of spatio-temporal similarity measure proposed by Hwang [4] and this issue is addressed by the authors of this paper in [7] by finding similarity percentage based on number of points a trajectory is visited, with applications in security informatics.

The first and foremost question needed to be considered in clustering web sessions is how to measure the similarity between two web sessions. Most of the previous related works apply either Euclidean distance for vector or set similarity measures, Cosine or Jaccard Coefficient. For example, a novel path clustering method was introduced [8] based on the similarity of the history of user navigation. A similar method [1] which first uses the longest common sub-sequences between two sessions through dynamic programming, and then the similarity between two sessions is defined through their relative time spent on the longest common sub-sequence. Methods have been developed [10] considering each session as a sequence and borrow the idea of sequence alignment in bioinformatics to measure similarity between sequences of page accesses. One of the recent work in web session clustering [2] uses dynamic programming technique in finding sequence similarity after finding structural similarity. In none of the papers the concept of spatio-temporal measure has been discussed in finding session similarity, which add the time component to the spatial similarity measure. There are resemblance of spatio-temporal similarity measure of network constrained trajectories and trajectories of user sessions in web usage mining [9] but there is no work in the literature exploring it in detail. This paper explore this idea proposing an algorithm for spatio-temporal similarity set formation using web access log data.

3 Trajectory Similarity of Moving Objects and Web User Sessions

Trajectory database management has emerged due to the profusion of mobile devices and positioning technologies like GPS or recently the RFID (Radio Frequency Identification). Studying people and vehicle movements within some road network is both interesting and useful especially if we could understand, manage and predict the traffic phenomenon. In essence, by studying the massive flow of traffic data we can monitor and discover traffic related patterns. There are several advantages in representing trajectory data in road network distance rather than Euclidian Distance [6]. The dimensionality reduction is one advantage where a trajectory is represented as a set of (loc, t) points instead of (x,y,t) points in Euclidian space.

3.1 Trajectory Similarity in Spatial Networks

Let T be a trajectory in a spatial network, represented as

$$T = ((b_1,t_1),(b_2,t_2),(b_3,t_3),\ldots\ldots\ldots\ldots,(b_n,t_n))$$

where n is the trajectory description length, bi denotes a location in binary string [6] and t_i is the time instance (expressed in time units, eg. seconds) that the moving object reached node bi, and $t_1 < t_i < t_m$, for each $1 < i < m$. It is assumed that moving from a node to another comes at a non-zero cost, since at least a small amount of time will be required for the transition.

The similarity types related to road network environment are

(i) Finding objects moving through certain Points of Interest
(ii) Finding Objects moving through Certain Times of Interest
(iii) Finding Objects moving through certain Points of Interest and Time of Interest.

A detailed discussion of these methods and evaluation can be found in [4] and modified algorithms in [7]. These methods and algorithms are the base line content of the section 3.2 where similarity of web user session trajectories and proposed spatio-temporal similarity algorithms are being discussed.

3.2 Similarity in Web User Session Trajectory

There are a number of resemblance between trajectories of moving objects on road network and trajectories of web click-streams. A web link structure can also assumed as network constrained since the user clicks can traverse only through predefined paths which is stored as web structure links in web storage space. By analyzing the URL path of each user, we are able to determine paths that are very similar, and therefore effective caching strategies can be applied.

3.2.1 Web User Session

The usage data is one that describes the pattern of usage of web pages like IP address, Page references, date and time of accesses. A web server log records the browsing behavior of site visitors and hence it is important source for usage information like Cookies and query data in separate logs. Cookies are tokens generated by the web server for individual client browsers in-order to automatically track the site visitors. Due to the stateless connection of HTTP protocol, tracking of individual users is not an easy task. A user is defined as the single individual accessing files from one or more web servers through a browser. A user stream session is the click-streams of page view for a single user across the entire web. In this work we consider a user session as a sequence of web clicks of individual user on a particular website.

3.2.2 Comparison of Network Constrained Moving Object Trajectory and Web User Session Trajectory

A moving object Trajectory is the path of a moving object along a road during a given period of time. A web user session trajectory is a set of user clicks during a period of time. Both are made up of individual nodes which denotes either location or url and time instant as shown in Fig 1.

Some of the variations in these two representations are

(i) A node in Moving Object Trajectory(MOT) is the name of a location on road where as a node in Web user session trajectory is a url name or web page name.

(ii)The spatial distance between two nodes in MOT is the network distance between the locations where as in the case of web user session trajectory it is the structural distance of the web page in the tree made by the link structure.

(iii) In the case of MOT the direction of object movement is bidirectional but in web user session trajectories it is unidirectional.

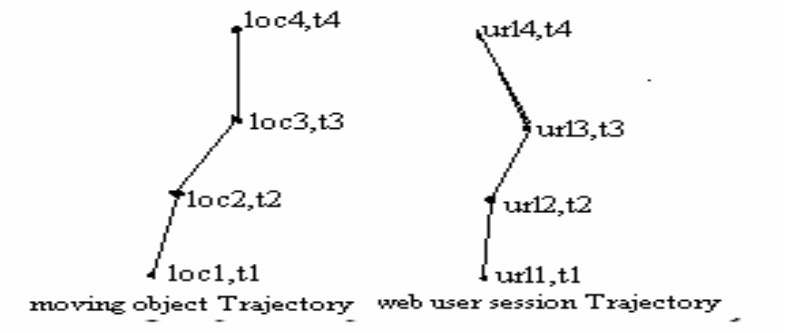

Fig. 1. Comparison of Moving Object Trajectory and Web User session Trajectory

(iv) The time between two locations in MOT is the travel time required to reach a location from the previous location which also depends on speed of movement. In the case of Web user session trajectory the time between two url's is time lag between receipt of request of previous url and the next url page. It consists of page load time and page view time as shown in Fig 2.

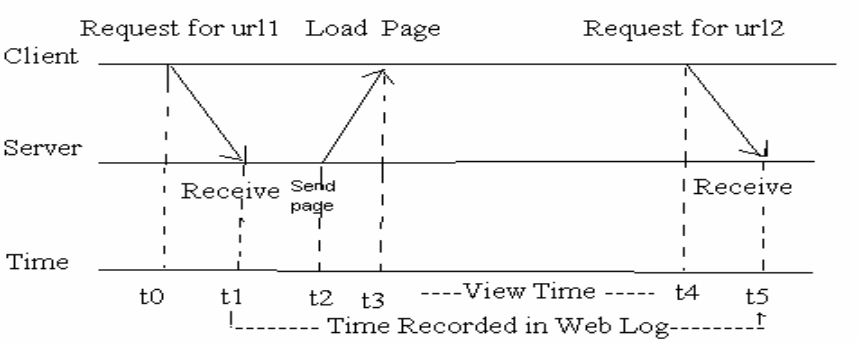

Fig. 2. Time of request between two url's

Though the web server records the time between two url's, as (t5 - t1), the actual page view time[2] of url1 will be (t4 - t3). In this paper, as we are establishing resemblance of web user session with moving object trajectory as shown in Fig 1, we consider the time recoded in web server log, instead of viewing time in finding the temporal similarity.

3.2.3 Finding Similar User Session in Web Access Logs

As discussed in the previous section, the path similarity of a user click-stream in web access log resembles the trajectory similarity problem of Moving object on Road Networks [7]. In this context Point of Interest (POI) is the interested Web URL's (We call URL's of Interest or UOI) chosen by the user as a query set. These are the URL's of specific interest to the user. For example in the case of Dark Web, which is a concept of web organization under terrorism, UOI's will be set of noted terrorism web

sites/pages in which the security people wants to find out how many people are visiting and the duration of such visits. Similarly Time of Interest (TOI) is the specific Time in which the user requesting the specified UOI's also has importance in practical situation. For example in web based security informatics one needs to find out similar viewing sessions based on a set of chosen pages visited and also based on viewing time in each page.

3.2.4 Definition of Web User Session Trajectory

The paper suggest a two step process of filtering trajectories based on spatial similarity and then refining similar trajectories based on temporal distance. For this the following definition is given, in comparison with definitions given in section 3.1.

Definition 1

Let S be a set of trajectories in a set of web user sessions, in which each trajectory is represented as

$$T = ((URL_1,t_1),(URL_2,t_2),(URL_3,t_3),\ldots\ldots\ldots\ldots,(URL_n,t_n))$$

where n is the trajectory description length, URL_i denotes a web page and ti is the time instance (expressed in time units, e.g. seconds) that the user requested for web page URL_i, and $t_1 < t_i < t_m$, for each $1 < i < m$. It is assumed that moving from a URL node to another comes at a non-zero cost, since at least a small amount of time will be required for loading the requested page by the server and viewing by the client.

3.2.5 Creation of User Session Trajectories from Web Access Logs

The goal of session identification is to divide the page accesses of each user into individual sessions. The following is the rules in identifying and distinguishing a user session from others from a web access log which we used in our experiment:

i) If there is a new user, there is a new session

ii) In one user session, if the referred page is null, there is a new session

iii) If the time between page requests exceeds a certain limit(30 minutes), it is assumed that the user is starting a new session, on the assumption that a user will generally may not spend an average of 30 minutes in a particular web page.

3.3 Finding Similar Web Usage Session Trajectories

We introduce the spatio-temporal similarity measures between two user-sessions, assuming resemblance to moving object trajectories in a constrained network. For moving object trajectories, finding spatio-temporal similarity [7] is a two step process of filtering trajectories based on spatial similarity and then refining similar trajectories based on temporal similarity. But in the case of web user sessions the spatial similarity is a combined measure of how many common URL's in both the sessions, structural similarity of web pages and the sequences of appearance of URL's in the trajectories. Therefore in this paper we are proposing a three stage process for spatio-temporal similarity of web user session trajectories as given in Fig 3.

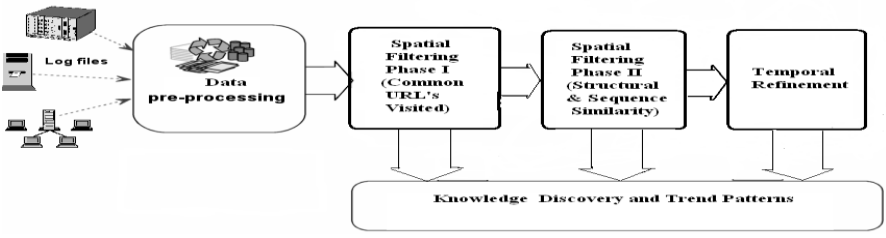

Fig. 3. Spatio-Temporal Similarity Framework

The practical importance of spatio-temporal similarity measure which includes the structural and sequence similarity is illustrated in the following case.

In Security Informatics concerning dark web environment the security officials wants to know when, how many and what sequence insurgents or suspicious people browsing through terrorist sites and based on which their common tactics can be judged.

3.3.1 Spatial Similarity Matrix Based on UOI

We introduce the similarity measures between two user-sessions, assuming both are two moving object trajectories in a constrained network [7] incorporating concepts discussed in [15]. Here the spatial similarity is measured based on three concepts.

a) Common URL's visited by two trajectories
b) Structural similarity of web pages in the trajectories
c) Sequence Similarity of user sessions.

a) Common URL's visited by two trajectories

Let T_i and T_j be two web user session trajectories. We introduce a Spatial Similarity measure $SSim_C(T_i,T_j)$ which attempt to incorporates number of URL's involved in trajectories. This similarity matrix measures the number of common URL's accessed during the two sessions relative to the total number of URL's accessed in both sessions.

Definition 2

Let T_i and T_j be two user session trajectories. The spatial similarity measure between these two trajectories is defined based on the concept that the trajectories passes through how many URL's in common.

$SSim_C(T_i,T_j)$ = Number of URL's common to Ti and Tj/Total Number of URL's in Ti and Tj.

Note that the above Similarity measure satisfies the following properties.

i. $SSim_C(T_i,T_j)$)is lies in the range [0,1]
ii. $SSim_C(T_i,T_j) = SSim_C(T_j,T_i)$

b) Structural Similarity

In order to measure Structural similarity[2], the content of pages is not considered but it is baed on the paths leading to a web page (or script). For example Suppose "/faculty/pub/journal/ieee.htm" and "faculty/pub/conference.htm" are two requested pages in web log. By assuming the page connectivity as a tree structure with root coming as home page , each level of a URL can be represented by a token from left to right in the order 0,1,2,3 etc. Thus, the token string of the full path of a URL is the concatenation of all the representative tokens of each level. We get different token strings of the URL correspond with the different requested pages traversing the tree structure of the web site. Assuming that the two token strings of the above two web pages are "0222" and "021" respectively, we compare each corresponding token of the two token strings one by one from the beginning until the first pair of tokens is different.

Marking Llonger as the larger value of (l_1, l_2) where l_1 and l_2 are lengths of the two token strings, we give weight to each level of the longer token: the last level is given weight 1, the second to the last level is given weight 2, the third to the last level is given weight 3, and so on, until the first level which is given weight Llonger. The similarity between two token strings is defined as the sum of the weights of those matching tokens divided by the sum of the total weights.

Definition 3

The Structural similarity of two web pages is defined as

$WSim_S (P_i,P_j)$ = Sum of the weights of matching token strings/Sum of the total weights.

For our example

P1= "/faculty/pub/journal/ieee.htm" and P2= "/faculty/pub/conference.htm"
Llonger = 4 ,
Token string in P1: 0 2 2 2
Token string in P2: 0 2 1
Weight of each token: 4 3 2 1

Since the first and second digit matches in P1 and P2, we take the weights of the matching locations 4 and 3 respectively.

The similarity of the two requested web pages is

$WSim_S (P1,P2)= (4 + 3)/(4+3+2+1) = 0.7.$

Note that the above Similarity measure satisfies the following properties.

i. $WSim_S (P_i,P_j) = WSim_S (P_j ,P_i)$
ii. $WSim_S (P_i,P_j)$ lies in the range [0,1]
iii. $WSim_S (P_i,P_j) = 0$ when there is no structural similarity between pages Pi and Pj
iv. $WSim_S (P_i,P_j) = 1$ when two pages Pi and Pj are exactly same.

c) Sequence Similarity of web sessions

We use a scoring system which helps to find the optimal matching between two session sequences. An optimal matching is an alignment with the highest score. The

score for the optimal matching is then used to calculate the similarity between two sessions. We find the optimal matching using dynamic programming techniques[2] to compute the similarity between web sessions as given in the following algorithm

---

Input: Pair of session sequences which constitutes each trajectories T_1,T_2
Output: Optimal Similarity score value OSSV

---

1. Matrix DPM is created with K+1 columns and N+1 rows where k and N corresponds to the sizes of T_1 and T_2 sequences respectively.
2. Align the sequences by providing gap between the sequence so that two sequences can be matched as much as possible.
3. Place sequences T1 in top of Matrix and sequences T2 on left side of the matrix.
4. Assign top left corner of matrix =0
5. For each pair of web pages /*Find optimal path */
6. Find $WSim_S(P_i,P_j)$
7. Find score = -10+30* $WSim_S(P_i,P_j)$
8. DPM cell entry=max(left entry+score,topentry+score,above left entry+score)
   /* Find path for every two sequence pair of web pages as follows */
9. If both the pages are matching
10. Diagonal move
11. Else if gap on top horizontal sequence
12. Right move
13. Else if gap on left vertical sequence
14. Down move
15. End if
16. Optimal Similarity score value(OSSV) = value at the lower right corner of the matrix DPM.
17. Return OSSV

After finding the final score for the optimal session alignment, the final similarity between sessions is computed by considering the final optimal score and the length of the two sessions. In our method, we first get the length of the shorter session lshorter, then the similarity between the two sessions is achieved through dividing the optimal matching score by 20* lshorter because the optimal score cannot be more than (20* lshorter) in our scoring system. Thus the spatial similarity measure between two web session trajectories is defined based on the concept of Sequence similarity as $Ssim_s(T_i,T_j)$ = Optimal score / (20* length of the shorter trajectory)

Note that the value of $Ssim_s(T_i,T_j)$ lies between 0 and 1.

3.3.2 Combined Spatial Similarity Matrix

The combined spatial similarity measure is obtained by

$Ssim_{cs}(T_i,T_j) = (SSim_C(T_i,T_j) + Ssim_s(T_i,T_j))/2$

$Ssim_{cs}(T_i,T_j)$ satisfies the following properties.

i. $Ssim_{cs}(T_i,T_j)$ is in [0,1]
ii. $Ssim_{cs}(T_i,T_j) = Ssim_{cs}(T_j,T_i)$

The combined spatial similarity measure is used as the distance measure in locating the objects within the similarity set.

3.3.3 Temporal Distance Measure Based on UOI and TOI

The similarity measure defined in the previous section takes into consideration only the spatial concept, which consists of structural similarity and sequence similarity. In real applications, the time information associated with each trajectory is also very important. So to measure the similarity we have to consider the concept of space and time together. Here we are considering the temporal distance by taking the difference in web page request time as shown in Fig 4.

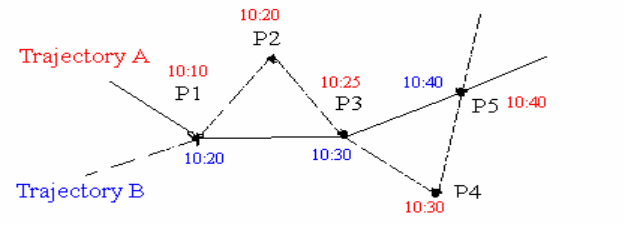

Fig. 4. Temporal Distance of Two Trajectories

Temporal distance between Trajectory A (T_A) and Trajectory B (T_B) will be $dist_T(T_A,T_B)$ = Differences in time at common URL's visited

$$P1,P3,P5 = 10+5+0 = 15$$

Definition 4

The temporal distance between two user session trajectories is defined as

$$dist_T(Ti,Tj) = \frac{1}{m*z} \sum |t(Ti.URL_k) - t(Tj.URL_k)|$$

where summation will go for only common URL's visited by both the trajectories(m) and z is a constant chosen by the user to denotes the time span in which the similarity is being measured. For example z is fixed to 24, when similarity is measuring within 24 hours in a day. Note that this distance measure also satisfies the following properties

i. $dist_T(Ti,Tj)$ is in [0,1]
ii. $dist_T(Ti,Tj) = dist_T(Tj,Ti)$

The difference between the above time similarity measure with the time similarity measure in [2] is that the later considering the viewing time similarity of a web page.

3.3.4 WEBTRSIM – Similarity Tool for Web User Session Trajectory

We propose a web session similarity set formation algorithm WEBTRASIM(Web user session Trajectory Similarity Algorithm). The objective of this algorithm is to form spatio-temporal set of trajectories, which are similar to the query trajectory formed from url's of interest(UOI) and Time of Interest (TOI).

Algorithm: To Create Spatio-Temporal Similarity Set based on UOI and TOI

Input: Input user session trajectories TR_{IN}, spatial threshold ρ, δ, temporal threshold-t σ , query trajectory tr_Q which is made up of UOI set and TOI set in required sequence; Output: Spatio-Temporal cluster set CTR containing trajectories similar to query trajectory tr_Q

---

1. Set $TR_{Candidate}$, Trout, CTR as Empty set
2. n1= number of URL's in tr_Q , , n2= number of URL's in TR_{IN},
3 For each t_r in TR_{IN},
4. n2= number of URL's in t_r , t_r.s=0
   /* Spatial filtering(first phase) based on common url's visited*/
5. For each u in tr_Q
6. If u is on tr then $t_r.s = t_r.s+1$
7. End For
8. $t_r.s= t_r.s/(n1+n2)$
9. If $t_r.s > \rho$ then $TR_{Candidate} = TR_{Candidate} \cup \{t_r\}$
10. End For
    /* Spatial filtering(second phase) based on structural and sequence similarity measure */
11. For each $t_r \in TR_{Candidate}$
12. n2=no of URL's in t_r
13. if n1<n2 then lshort=n1 else lshort=n2
14. Opscore= SSCORE-DPROG(t_r,tr_q)
15. Ssims = opscore/ (20*lshort) , tr.s=(tr,s+ Ssims)/2
16. If Ssims > δ then $TR_{OUT} = TR_{OUT} \cup \{ t_r \}$
17. End For
    /* Temporal refinement based on temporal distance */
18. For each tr in TR_{OUT},
19. n2= number of URL's in t_r , t_r.t=0, m=0
20. For each u in tr_Q
21. If u is on t_r then
22 $t_r.t = t_r.t + | t_r.u.t - tr_Q.u.t |$
23 m=m+1
24. End For
25. $dist_T(t_r, tr_Q) = 1/(m*24*60)* t_r.t$ /* one day- 24hours as the time span*/
26. $tr._{temp\text{-}dist} = dist_T(t_r, tr_Q)$
27 if $dist_T(t_r, tr_Q) < \sigma$ then $CTR = CTR \cup \{ t_r \}$
28 .End For
29 .return CTR

The algorithm will make a similarity set based on the spatio-temporal measure discussed above. One advantage of the method is that along with each trajectory in the set, spatial and temporal similarity measures(t_r.s, $tr._{temp\text{-}dist}$) are also stored which will be used later to visualize the scattering of similar user sessions within the similarity

set, useful for finding nearest neighbor trajectories based on a given UOI and TOI. A major difference of WEBTRASIM algorithm with the algorithm – WSCBIS proposed in [2] is that the later will take temporal similarity measure only when two trajectories are spatially similar. But our approach take a combined similarity measure since spatio-temporal features have to be considered in finding more accurate similarity.

4 Experimental Evaluation

To validate the efficiency of the clustering algorithm discussed above, we have made an experiment with the web server log publically available with statistics about raw data is shown in Table 1. The initial data source of our experiment is for one day on 26th Aug 2009, in which the size is 5372MB. The only cleaning step performed on this data was the removal of references to auxiliary files (image files)

Table 1. Summary of access log characteristics(Raw Data)

Item	Count
Total Requests(No of entries)	47,685
Average Request/Hour	1986.875
Total Bytes Transferred MB	304198.707
Average Bytes transferred per hour	12167.94
No of requests after pre-processing	33,864
No of Users	1,907
No of user sessions	1,986

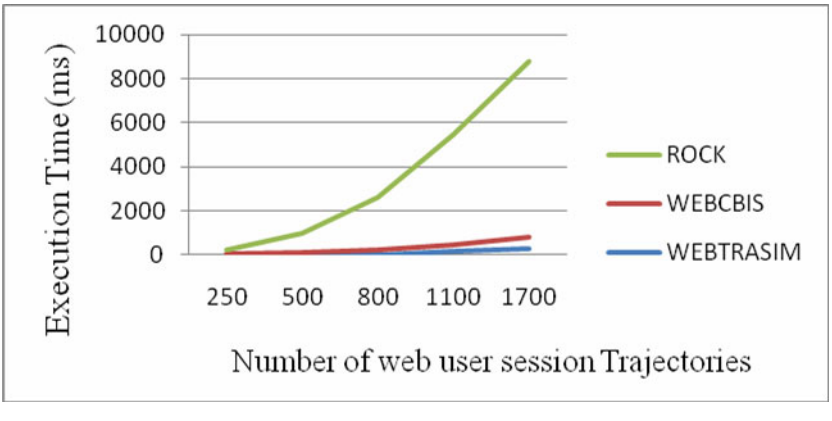

Fig. 5. Comparison of Execution Time

Our experiments were performed on a 2.8GHz Core-2 Duo Processor, 1GB of main memory, Windows XP professional using the visual Programming Language Visual Basic. As shown in above Table, after data cleaning, the number of requests declined from 47,685 to 33,864. We have applied the algorithms WSCBIS, ROCK and WEBTRASIM on the number of user sessions obtained as shown in Table 1, by giving a UOI consists of 4 urls with ρ=0.95, δ=0.95 and σ=0.05,.

Fig 5 shows that the execution time linearly increase when the number of web user session increases and the performance of WEBTRASIM is almost similar to WSCBIS even when it extends temporal similarity measure in addition to spatial similarity. Thus we claim that WEBTRASIM could become a better tool for spatio-temporal similarity of web user session trajectories.

5 Conclusion

The WWW, an effective information presentation and dissemination tool, has been widely used by terrorist groups as a communication medium. The Web presence of these terrorist groups reflects their different characteristics and may provide information about planned terrorist activities. Thus, monitoring and studying the content, structural characteristics and usage pattern of terrorist websites may help us to analyze and even to predict the activities of terrorist groups. In this paper we have proposed a method to measure the spatio-temporal similarity of web user sessions, by introducing the algorithm WEBTRASIM. Our experiments with a web access log shows that the algorithm performs equally well with similar algorithms even with the extension of temporal similarity measure. We are planning to use this similarity measure for web user session trajectory clustering to extract various browsing patterns to test with different dark web access logs.

Acknowledgements

This research is being supported by University Grants Commission(UGC) under Ministry of Human Resources, Government of India under the scheme of Grant to Academic Research for Faculty members in Universities vide order No. 37-633/2009 (SR).

References

1. Banerjee, G.A.: Clickstream clustering using weighted longest common subsequences. In: Proc. of the Workshop on Web Mining, SIAM Conference on Data Mining, Chicago, pp. 158–172 (2009)
2. Chaofeng, L.: Research on Web Session Clustering. Journal of Software 4(5) (2009)
3. Weimann, G.: How Modern Terrorism Uses the Internet, United States Institute of Peace, Special Report 116 (2004), `http://www.terror.net`
4. Hwang, J.-R., Kang, H.-Y., Li, K.-J.: Searching for Similar Trajectories on Road Networks Using Spatio-temporal Similarity. In: Manolopoulos, Y., Pokorný, J., Sellis, T.K. (eds.) ADBIS 2006. LNCS, vol. 4152, pp. 282–295. Springer, Heidelberg (2006)
5. Lee, E., Leets, L.: Persuasive storytelling by hate groups online - Examining its effects on adolescents. American Behavioral Scientist 45, 927–957 (2002)

6. Abraham, S., Lal, P.S.: Trigger based security alarming scheme for moving objects on road networks. In: Yang, C.C., Chen, H., Chau, M., Chang, K., Lang, S.-D., Chen, P.S., Hsieh, R., Zeng, D., Wang, F.-Y., Carley, K.M., Mao, W., Zhan, J. (eds.) ISI Workshops 2008. LNCS, vol. 5075, pp. 92–101. Springer, Heidelberg (2008)
7. Abraham, S., Lal, P.S.: Trajectory Similarity of Network Constrained Moving Objects and Applications to Traffic Security. In: Chen, H., Chau, M., Li, S.-h., Urs, S., Srinivasa, S., Wang, G.A. (eds.) PAISI 2010. LNCS, vol. 6122, pp. 31–43. Springer, Heidelberg (2010)
8. Shahabi, C., Zarkesh, A., Adibi, J.: Knowledge discovery from users' web-page navigation. In: Proceedings of the 7th International Workshop on Research Issues in Data Engineering (RIDE 1997) High Performance Database Management for Large-Scale Applications, pp. 20–31. IEEE Computer Society, Washington, DC, USA (1997)
9. Tiakas, E.: Searching for similar trajectories in spatial Networks. J. System Are (2009), doi:10.1016/j.jss.2008.11.832Y
10. Wang, W., Zaane, O.R.: Clustering Web sessions by sequence alignment. In: Proceedings of the 13th International Workshop on Database and Expert Systems Applications, pp. 394–398. IEEE Computer Society, Washington, DC (2002)
11. Chen, H., Chung, W., Xu, J., Wang, G., Qin, Y., Chau, M.: Crime data mining: A general framework and some examples. Computer 37, 50–54 (2004)
12. Xu, J., Chen, H., Zhou, Y., Qin, J.: On the Topology of the Dark Web of Terrorist Groups. In: Mehrotra, S., Zeng, D.D., Chen, H., Thuraisingham, B., Wang, F.-Y. (eds.) ISI 2006. LNCS, vol. 3975, pp. 367–376. Springer-Verlag, Heidelberg (2006)