ISSN: 2229-6956(ONLINE)
DOI: 10.21917/ijsc.2012.0064

DERIVING USER ACCESS PATTERNS AND MINING WEB COMMUNITY WITH WEB-LOG DATA FOR PREDICTING USER SESSIONS WITH PAJEK

S. Balaji[1] and S. Sasikala[2]
[1]*Department of Computer Science, Sengunthar College of Engineering, India*
E-mail: hereiambalaji@yahoo.com
[2]*Department of Computer Science, K.S.R College of Arts and Science, India*
E-mail: sasi_sss123@rediff.com

Abstract
Web logs are a young and dynamic media type. Due to the intrinsic relationship among Web objects and the deficiency of a uniform schema of web documents, Web community mining has become significant area for Web data management and analysis. The research of Web communities extents a number of research domains. In this paper an ontological model has been present with some recent studies on this topic, which cover finding relevant Web pages based on linkage information, discovering user access patterns through analyzing Web log files from Web data. A simulation has been created with the academic website crawled data. The simulation is done in JAVA and ORACLE environment. Results show that prediction of user session could give us plenty of vital information for the Business Intelligence. Search Engine Optimization could also use these potential results which are discussed in the paper in detail.

Keywords:
Web Log, Mining, Clustering, Social Network Analysis

1. INTRODUCTION

The analysis of Web log files may give information that are useful for improving the services offered by Web portals and information access and retrieval tools, giving information on problems occurred to the users.

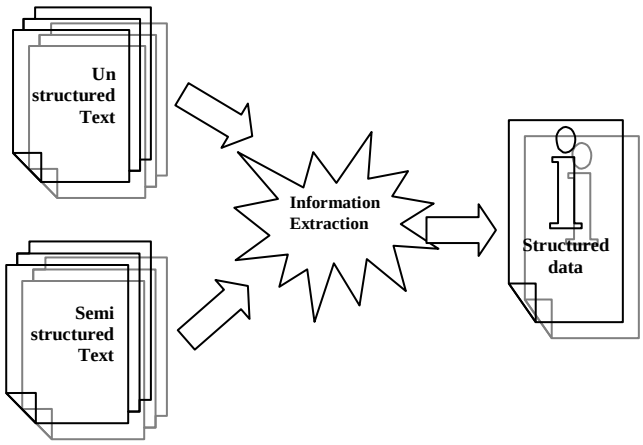

Fig.1. Information extraction from web 2.0

2. WORKING SCENARIO

In this paper, we aimed to investigate employing web clustering for web community mining and analysis in terms of web page communities, web user access patterns and co-clusters of web pages and uses as well. The extraction of the data from Web logs gives access to information that have to be managed efficiently in order to be able to exploit them for analyses.

3. INFORMATION EXTRACTION FROM WEBLOGS AND COMMUNITIES

Many communities, either in an explicit or implicit form, have existed in the Web today, and their number is growing at a very fast speed. Discovering communities from a network environment such as the Web has become an interesting research problem recently. Network structures like the Web can be abstracted into directional or non-directional graphs with nodes and links.

It is usually rather difficult to understand a network's nature directly from its graph structure, particularly when it is a large scale complex graph. Data mining is a method to discover the hidden patterns and knowledge from a huge network. The mined knowledge could provide a higher logical view and more precise insight of the nature of a network, and will also dramatically decrease the dimensionality when trying to analyze the structure and evolution of the network. Information extraction from blogs and web2.0 scenario would be mentioned in Fig.1.

4. RELATED WORK

Quite a lot of work has been done in mining the implicit communities of users, web pages or scientific literature from the Web or document citation database using content or link analysis [4, 5, 6, and 7]. Several different definitions of community were also raised in the literature. In [5], a web community is a number of representative authority web pages linked by important hub pages that share a common topic as shown in Fig.2(a). In [6], a web community is a highly linked bipartite sub-graph and has at least one core containing complete bipartite sub graph as shown in Fig.2(b). In [4], a set of web pages that linked more pages in the community than those outside of the community could be defined as a web community (see Fig.2(c)). Also, a research community could be based on a single most-cited paper and contain all papers that cite it [7] (see Fig.2(d)).

5. COMMUNITY MINING

Community mining allows us to find clusters of people who are densely connected to each other, but only sparsely connected to the rest of the network. In a similar way that the formal organization charts break down the employees into smaller substructures, community mining finds natural communities in

the informal social network. Members of a community tend to mainly communicate with the other members of that community, and less so with people in the rest of the network. For instance, a community can represent members of three different engineering departments who are currently cooperating on the same project. Various scientific fields have developed a wide variety of algorithms for detecting communities in social networks. These algorithms mainly differ on how they define dense connections and the heuristics they employ to identify the dense clusters. Here we mined the academic website in which different department details are used for the experiment.

6. PROPOSED MODEL AND EXPERIMENTAL SETUP

Currently, the proposed system is applied to an academic society because researchers have various social relationships (e.g., from a student to a professor, from a company to a university) through their activities such as meetings, projects, and conferences. Fig.2 shows a sample community formed based on writers details. The tool used for the simulation is PAJEK.

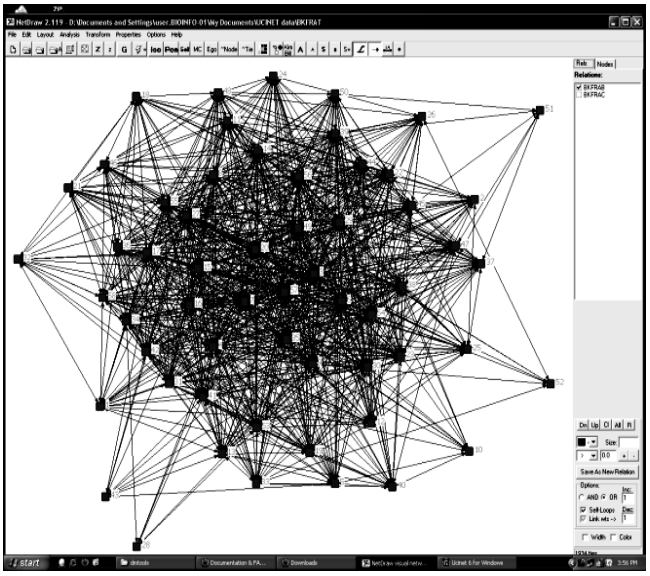

Fig.2. Pajek Rendition of the Community

Network analysts study the patterning of the social connections that link sets of actors. For the most part they seek to uncover either or both of two kinds of patterns. They often look for social groups collections of actors who are closely linked to one another. Or, alternatively, they look for social positions sets of actors who are linked into the total social system in similar ways. Some parameters that could be used in network analysis are explained in brief below,

Distance centrality of a vertex (D_c): proximity to the rest of vertices in the network. It is also called closeness centrality: the higher its value, the closer that vertex is to the others (on average). Given a vertex v and a graph G, it can be defined as,

$$D_C(v) = \frac{1}{\sum_{t \in G} d_G(v,t)} \quad (1)$$

$$B_C(v) = \sum_{s \neq v \neq t / in G} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (2)$$

Clustering coefficient of a vertex: The clustering coefficient c of a vertex measures the connectivity of its direct neighborhood. Given a vertex v in a graph G, it can be defined as the probability that any two neighbors of v be connected. Hence,

$$c(v) = \frac{E(v)}{k_v(k_v - 1)} \quad (3)$$

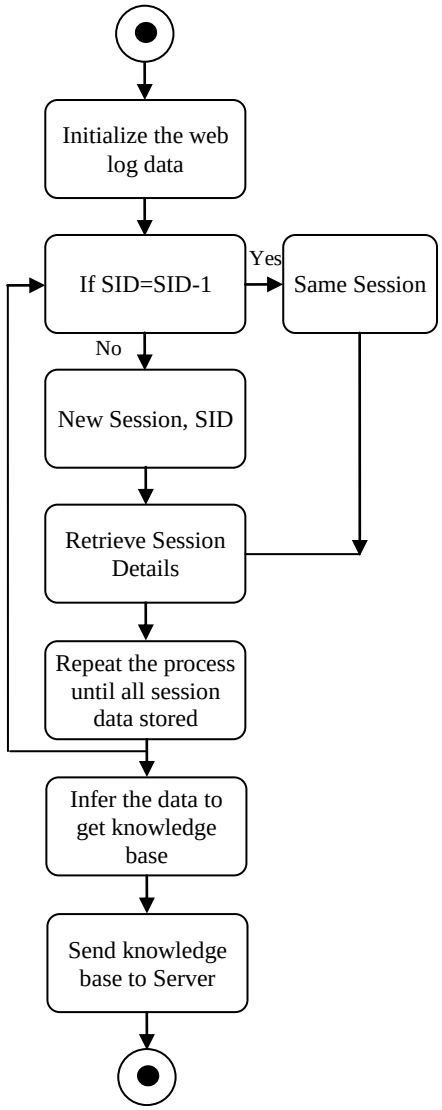

Fig.3. Session identification from user navigation

Weighted clustering coefficient of a vertex: The weighted clustering coefficient c_w of a vertex is an attempt to generalize the concept of clustering coefficient to weighted networks. Given a vertex v in a weighted graph G it can be defined as,

$$c_W(v) = \sum_{i \neq j \in N_G(v)} W_{ij} \frac{1}{k_v(k_v - 1)} \quad (4)$$

These features were assessed with the PAJEK tool and the considered data is the academic community in which the above defined features were retrieved. The simulated results are shown in Fig.3. Features were again fed into the JAVA interface

developed. Further analysis is done with the interface and user sessions are retrieved.

An algorithm about how to use (*m*-1)-item sets to generate m-item sets orderly is described in [3]. Below is the pseudo code of finding frequent item sets.

Phase I: Classifier

Generate 1-itemsets IS1 with minimal support S

k _ 2

***while** k*

Generate k-itemsets ISk using (k-1)-itemsets IS(k-1) with S

Prune IS (k-1) using ISk

k _ k +1

end

Put IS1 to ISm to itemsets set IS

Phase II:B-SIGNET

***for** every itemset I in IS*

Put objects in I to community C

do

Add objects not in C but having links to objects in C to C

Calculate ranking value of new added objects

***until** No more objects could be added*

Put a copy of C to communities set CS

Clear C

End

After analyzing the website with PAJEK the program is written to retrieve the sessions from the website. The time user interacts with the system or application is session. The user session starts when the user accesses the application and ends when the user quits the application.

Every user will have a unique IP address and assigned session ID. Website traffic is measured in terms of number of users using in one particular time in a site. A time slice may be assigned by the ADMIN. If user comes back within the time slice it will be treated as same session, since any number of visits within the given time slice will be treated as one session. If user re-enters after the time slice then it will be treated as a new session. Fig.4 depicts the session identification flow diagram for this paper. For a given user, session is based on the time slice value. When user re-enters the session will either continue or expired based on time slice.

The experimental evaluation was conducted using academic website Log file. The only cleaning step performed on this data was the removal of references to auxiliary files (e.g., image files). No other cleaning or preprocessing has been performed in the first phase. All evaluation tests were run on a dual processor Intel CPU 1.8 GHz Pentium 4 with 2 GB of RAM, operating system Windows XP. Our implementations run on JAVA and ORACLE. The results were given in Fig.5. Initially the data population process is carried out with the help of the crawler. Then the log data list is acquired. The session details are retrieved with the help of the log data. All are listed in Fig.5.

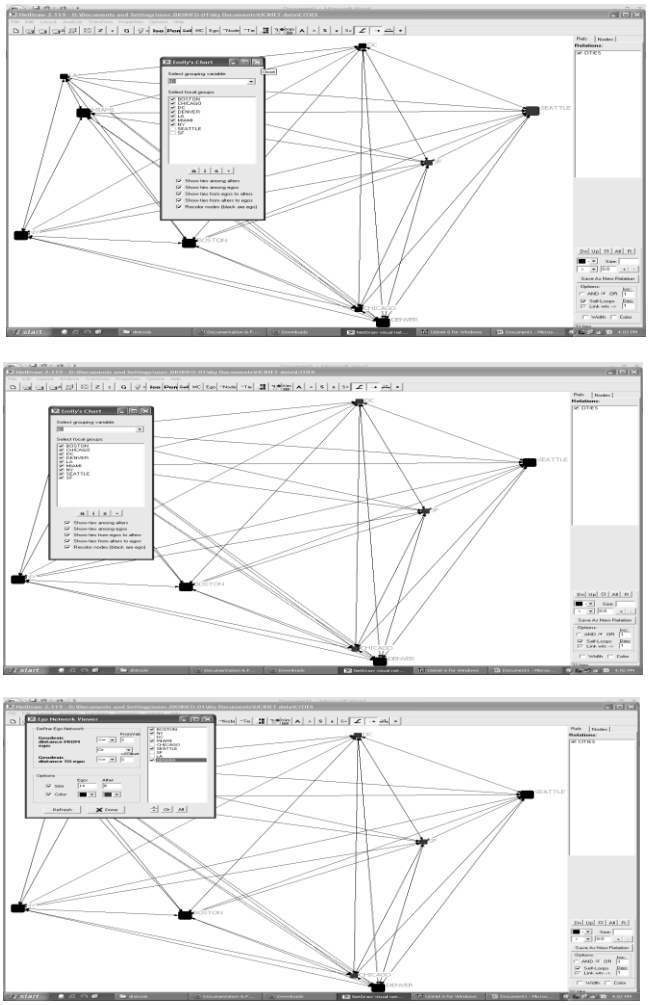

Fig.4. Simulated results for the Academic community feature extraction

7. RESULTS AND DISCUSSION

A sample result of the simulation has been offered in Fig. 6 which is created in open source web mining tool PAJEK [5]. The ranking function that is evaluated experimentally is based on the TF-IDF cosine similarity between the context of the query and the text of the entity's article:

$$score(q,e) = \cos(q.T, e.T) = \frac{q.T}{\|q.T\|}\frac{e.T}{\|e.T\|} \quad (5)$$

The factors $q.T$ and $e.T$ are represented in the standard vector space model, where each component corresponds to a term in the vocabulary, and the term weight is the standard TF-IDF score. The experimental evaluation was conducted using academic website Log file. The only cleaning step performed on this data was the removal of references to auxiliary files (e.g., image files). No other cleaning or preprocessing has been performed in the first phase. All evaluation tests were run on a dual processor Intel CPU 1.8 GHz Pentium 4 with 2 GB of RAM, operating system Windows XP. Our implementations run on JAVA and ORACLE. The results were given in Fig.5. Initially the data population process is carried out with the help of the crawler. Then the log data list is acquired. The session details are

retrieved with the help of the log data. All are listed in Fig.4. The WebPages used in the experiments are crawled from academic sites. In order to better show the effectiveness of the proposed framework, it is only selected some WebPages containing multiple mentions of the same entity. It is randomly sampled 25 pages for training and 100 pages for testing. With the help of retrieved session details, sessions/day (hour) is evaluated. It is given in Fig.5.

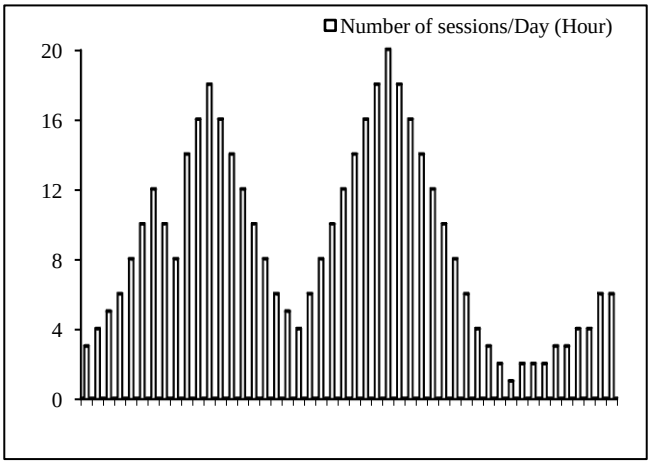

Fig.5. Evaluation of sessions/day/hour

Table.1. Session and User visiting pattern

SESSION SID-UID	WEB PAGES VISITED WPN – N, Web Page Number
S1-U1	WP1,WP3,WP5,WP1,WP15,WP9
S1-U2	WP1,WP6,WP14,WP4
S1-U3	WP10,WP2,WP11,WP7,WP13,WP6,WP3
S1-U4	WP11,WP7,WP13,WP1,WP5
S1-U5	WP9,WP8, WP11,WP7,WP13
S1-U6	WP3,WP8,WP12
S1-U7	WP8,WP4,WP12
S1-U8	WP5,WP1,WP15,WP9
S1-U9	WP1,WP7,WP4,WP9,WP10,WP11
S1-U10	WP1,WP8,WP5,WP1,WP15

With the help of retrieved session details, sessions/day (hour) and session/hour are evaluated. It is given in Fig.5, Fig.7 and Fig.8. Fig.6 is developed system.

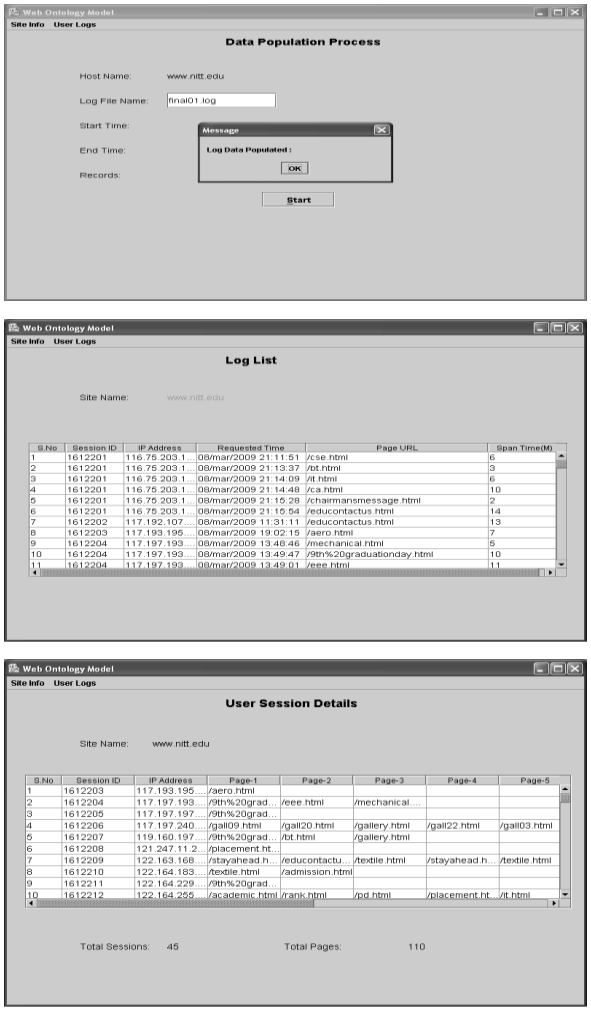

Fig.6. Screenshots of the JAVA interface of log-file analysis

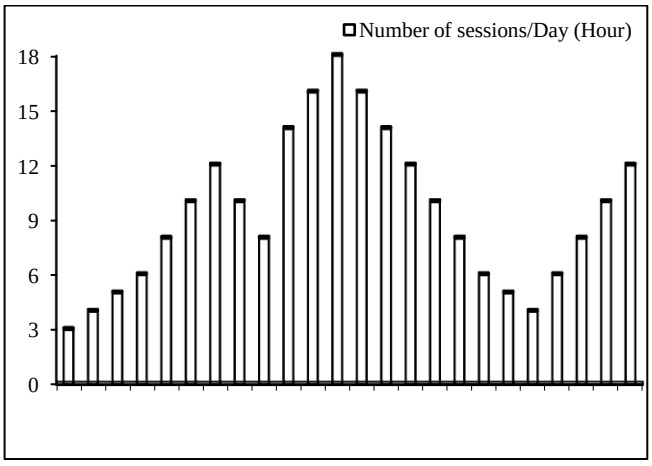

Fig.7. Evaluation of sessions/day (hour)

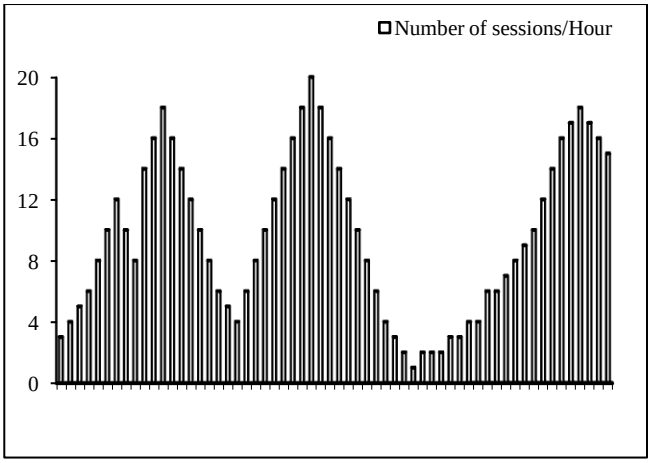

Fig.8. Evaluation of sessions/hour

In Fig.6 the developed system is given and initially the data is populated based on the user visiting pattern. Later the log list will be displayed retrieved from the user navigational pattern. Later the sessions will be identified for the community prediction. The average sessions per day and per hour are portrayed in Fig.7 and Fig.8. It shows the retrieved sessions.

8. CONCLUSION

This paper proposes a method to find the user sessions from the log data. This can be useful in SEO, website advertisements and online recommendations. This data can also be used for the recommender systems and web personalization. To retrieve the session details we have created and interface which uses the web log data and it shows a good performance. By predicting the active sessions the web mining could apply variety of techniques to utilize the users time. We also plan to analyze the evolution of communities. Since deriving graph communities to lay down the relationship could yield greater benefits.

APPENDIX – A

SAMPLE DATASET

1612201 116.75.203.100 08/mar/2009 21:08:29 /mechanical.html

101612201 116.75.203.100 08/mar/2009 21:10:29 /ece.html

151612201 116.75.203.100 08/mar/2009 21:11:51 /cse.html 6

1612201 116.75.203.100 08/mar/2009 21:13:37 /bt.html 3

1612201 116.75.203.100 08/mar/2009 21:14:09 /it.html 6

1612201 116.75.203.100 08/mar/2009 21:14:48 /ca.html 10

1612201 116.75.203.100 08/mar/2009 21:15:28 /chairmansmessage.html

1612201 116.75.203.100 08/mar/2009 21:15:54 /educontactus.html

1612202 117.192.107.187 08/mar/2009 11:31:11 /educontactus.html 13

1612203 117.193.195.232 08/mar/2009 19:02:15 /aero.html 7

1612204 117.197.193.113 08/mar/2009 13:48:46 /mechanical.html

1612204117.197.193.11308/mar/200913:49:47 9th%20graduationday.html

1612204 117.197.193.113 08/mar/2009 13:49:01 /eee.html 11

1612205117.197.197.20408/mar/200923:52:53 9th%20graduationday.html

1612206 117.197.240.83 08/mar/2009 11:26:44 /gall08.html 1

1612206 117.197.240.83 08/mar/2009 11:24:23 /gallery.html 13

REFERENCES

[1] L. Efimova and S. Fiedler, "Learning webs: Learning in weblog networks", *Proceedings of the IADIS International Conference Web Based Communities*, pp. 490-494, 2004.

[2] S. Balaji and S. Sasikala, "TD-Signet: Community Mining With Wsd Based On Implied Graph Structure In Social Networks", *International Journal of Engineering Science and Technology*, Vol. 3, No. 6, pp. 4588-4596, 2011.

[3] Mukul Joshi and Nikhil Belsare, "BlogHarvest: Blog Mining and Search Framework", *International Conference on Management of Data*, 2006.

[4] S. Balaji and S. Sasikala, "Signet: Web Information Retrieval with NE Disambiguation based on HMM and CRF", *International Journal of Machine Learning and Computing*, Vol. 2, No. 4, pp. 443-445, 2012.

[5] pajek.imfm.si/doku.php?id=download.