International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-5, March 2019

Efficient Approach for Weblog Analysis based on Maximum Frequency

Dharmendra Dangi, Amit Bhagat, Brijesh Bakariya

Abstract*: *Internet provides various services where a person interacts with each other. When a person performs any activity by internet then all the records stored on a web server. The data stored on the server called weblog data. This weblog contain lots of information about users. Now every person can get any information on a click. The huge amount of information stored on server. If we want to get the desired information from web server then it has to use some data mining techniques. Frequent pattern mining is one of the techniques for getting patterns from weblog. In this paper proposed an algorithm and framework for Pattern Analysis based on Maximum Frequency of Weblog (PAMFW) and also proposed a framework for pattern analysis.

Index Terms*: *Data Mining, Internet, Pattern Analysis, Web Server, Weblog.

Web log data are stored on the web server. If user interact with an internet then this log are stored. [1] [2]. Data preprocessing is also mandatory approach for preprocessing a server log. After preprocessing technique we can apply some data mining technique for getting patterns from weblog. This section described the proposed algorithm [3] [4]. Here algorithm is implemented for preprocessing and patterns generation from weblog. We have taken a dataset from ITA weblog repository [8]. There are following parameters of weblog which we are describing below [2] [7]. The data preprocessing like collection of data. It is a procedure to convert the various kinds of information such as text, image, audio, video etc. These kinds of method convert our web data into this preprocessing covert the data into a unique format [6] [9] [10]. Here we are taking a parameter time from log data [5][7]. In Table 1 the highest time durations are 4 hours. There are following descriptions of parameter taken from weblog IP Address Here we are considering IP address as a web user {wu1, wu2, wu3..... wun} is the set of web users. Time The time T = {time1, time2, time3.....timen} is the set of timeslots where web pages is visited by the user. URL The URL U {url1, url2, url3 .....urln} is the set of URL's for getting an information about the contents. Maximum Frequency (MF) An itemset X is called MF of X (Maximum Frequency) MF(X) >= TV (Threshold Value) otherwise those itemsets are discarded. There are following transaction table.

```
199.72.81.55 - - [01/Jul/1995:00:00:01 -0400] "GET /history/apollo/ HTTP/1.0" 200 6245
unicomp6.unicomp.net - - [01/Jul/1995:00:00:06 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985
199.120.110.21 - - [01/Jul/1995:00:00:09 -0400] "GET /shuttle/missions/sts-73/mission-sts-73.html HTTP/1.0" 200 4085
burger.letters.com - - [01/Jul/1995:00:00:11 -0400] "GET /shuttle/countdown/liftoff.html HTTP/1.0" 304 0
199.120.110.21 - - [01/Jul/1995:00:00:11 -0400] "GET /shuttle/missions/sts-73/sts-73-patch-small.gif HTTP/1.0" 200 4179
burger.letters.com - - [01/Jul/1995:00:00:12 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 304 0
burger.letters.com - - [01/Jul/1995:00:00:12 -0400] "GET /shuttle/countdown/video/livevideo.gif HTTP/1.0" 200 0
205.212.115.106 - - [01/Jul/1995:00:00:12 -0400] "GET /shuttle/countdown/countdown.html HTTP/1.0" 200 3985
d104.aa.net - - [01/Jul/1995:00:00:13 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985
129.94.144.152 - - [01/Jul/1995:00:00:13 -0400] "GET / HTTP/1.0" 200 7074
unicomp6.unicomp.net - - [01/Jul/1995:00:00:14 -0400] "GET /shuttle/countdown/count.gif HTTP/1.0" 200 40310
unicomp6.unicomp.net - - [01/Jul/1995:00:00:14 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786
unicomp6.unicomp.net - - [01/Jul/1995:00:00:14 -0400] "GET /images/KSC-logosmall.gif HTTP/1.0" 200 1204
d104.aa.net - - [01/Jul/1995:00:00:15 -0400] "GET /shuttle/countdown/count.gif HTTP/1.0" 200 40310
d104.aa.net - - [01/Jul/1995:00:00:15 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786
d104.aa.net - - [01/Jul/1995:00:00:15 -0400] "GET /images/KSC-logosmall.gif HTTP/1.0" 200 1204
129.94.144.152 - - [01/Jul/1995:00:00:17 -0400] "GET /images/ksclogo-medium.gif HTTP/1.0" 304 0
199.120.110.21 - - [01/Jul/1995:00:00:17 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1713
ppptky391.asahi-net.or.jp - - [01/Jul/1995:00:00:18 -0400] "GET /facts/about_ksc.html HTTP/1.0" 200 3977
net-1-141.eden.com - - [01/Jul/1995:00:00:19 -0400] "GET /shuttle/missions/sts-71/images/KSC-95EC-0916.jpg HTTP/1.0" 200 34029
ppptky391.asahi-net.or.jp - - [01/Jul/1995:00:00:19 -0400] "GET /images/launchpalms-small.gif HTTP/1.0" 200 11473
205.189.154.54 - - [01/Jul/1995:00:00:24 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985
waters-gw.starway.net.au - - [01/Jul/1995:00:00:25 -0400] "GET /shuttle/missions/51-l/mission-51-l.html HTTP/1.0" 200 6723
ppp-mia-30.shadow.net - - [01/Jul/1995:00:00:27 -0400] "GET / HTTP/1.0" 200 7074
205.189.154.54 - - [01/Jul/1995:00:00:29 -0400] "GET /shuttle/countdown/count.gif HTTP/1.0" 200 40310
alyssa.prodigy.com - - [01/Jul/1995:00:00:33 -0400] "GET /shuttle/missions/sts-71/sts-71-patch-small.gif HTTP/1.0" 200 12054
ppp-mia-30.shadow.net - - [01/Jul/1995:00:00:35 -0400] "GET /images/ksclogo-medium.gif HTTP/1.0" 200 5866
dial22.lloyd.com - - [01/Jul/1995:00:00:37 -0400] "GET /shuttle/missions/sts-71/images/KSC-95EC-0613.jpg HTTP/1.0" 200 61716
smyth-pc.moorecap.com - - [01/Jul/1995:00:00:38 -0400] "GET /history/apollo/apollo-13/images/70mC314.GIF HTTP/1.0" 200 101267
205.189.154.54 - - [01/Jul/1995:00:00:40 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786
ix-or12-01.ix.netcom.com - - [01/Jul/1995:00:00:41 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985
ppp-mia-30.shadow.net - - [01/Jul/1995:00:00:41 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786
ppp-mia-30.shadow.net - - [01/Jul/1995:00:00:41 -0400] "GET /images/MOSAIC-logosmall.gif HTTP/1.0" 200 363
205.189.154.54 - - [01/Jul/1995:00:00:41 -0400] "GET /images/KSC-logosmall.gif HTTP/1.0" 200 1204
ppp-mia-30.shadow.net - - [01/Jul/1995:00:00:41 -0400] "GET /images/USA-logosmall.gif HTTP/1.0" 200 234
ppp-mia-30.shadow.net - - [01/Jul/1995:00:00:43 -0400] "GET /images/WORLD-logosmall.gif HTTP/1.0" 200 669
ix-or12-01.ix.netcom.com - - [01/Jul/1995:00:00:44 -0400] "GET /shuttle/countdown/count.gif HTTP/1.0" 200 40310
gayle-gaston.tenet.edu - - [01/Jul/1995:00:00:50 -0400] "GET /shuttle/missions/sts-71/mission-sts-71.html HTTP/1.0" 200 12040
piweba3y.prodigy.com - - [01/Jul/1995:00:00:54 -0400] "GET /shuttle/missions/sts-71/sts-71-patch-
```

Fig. 1 Unprocessed Data

199.72.81.55	[01/Jul/1995:00:00:01	GET /history/apollo/ HTTP/1.0	200	6245
unicomp6.unicomp.net	[01/Jul/1995:00:00:06	GET /shuttle/countdown/ HTTP/1.0	200	3985
199.120.110.21	[01/Jul/1995:00:00:09	GET /shuttle/missions/sts-73/mission-sts-73.html HTTP/1.0	200	4085
burger.letters.com	[01/Jul/1995:00:00:11	GET /shuttle/countdown/liftoff.html HTTP/1.0	304	0
199.120.110.21	[01/Jul/1995:00:00:11	GET /shuttle/missions/sts-73/sts-73-patch-small.gif HTTP/1.0	200	4179
burger.letters.com	[01/Jul/1995:00:00:12	GET /images/NASA-logosmall.gif HTTP/1.0	304	0
burger.letters.com	[01/Jul/1995:00:00:12	GET /shuttle/countdown/video/livevideo.gif HTTP/1.0	200	0
205.212.115.106	[01/Jul/1995:00:00:12	GET /shuttle/countdown/countdown.html HTTP/1.0	200	3985
d104.aa.net	[01/Jul/1995:00:00:13	GET /shuttle/countdown/ HTTP/1.0	200	3985
129.94.144.152	[01/Jul/1995:00:00:13	GET / HTTP/1.0	200	7074
unicomp6.unicomp.net	[01/Jul/1995:00:00:14	GET /shuttle/countdown/count.gif HTTP/1.0	200	40310
unicomp6.unicomp.net	[01/Jul/1995:00:00:14	GET /images/NASA-logosmall.gif HTTP/1.0	200	786
unicomp6.unicomp.net	[01/Jul/1995:00:00:14	GET /images/KSC-logosmall.gif HTTP/1.0	200	1204
d104.aa.net	[01/Jul/1995:00:00:15	GET /shuttle/countdown/count.gif HTTP/1.0	200	40310
d104.aa.net	[01/Jul/1995:00:00:15	GET /images/NASA-logosmall.gif HTTP/1.0	200	786
d104.aa.net	[01/Jul/1995:00:00:15	GET /images/KSC-logosmall.gif HTTP/1.0	200	1204
129.94.144.152	[01/Jul/1995:00:00:17	GET /images/ksclogo-medium.gif HTTP/1.0	304	0
199.120.110.21	[01/Jul/1995:00:00:17	GET /images/launch-logo.gif HTTP/1.0	200	1713
ppptky391.asahi-net.or.jp	[01/Jul/1995:00:00:18	GET /facts/about_ksc.html HTTP/1.0	200	3977
net-1-141.eden.com	[01/Jul/1995:00:00:19	GET /shuttle/missions/sts-71/images/KSC-95EC-0916.jpg HTTP/1.0	200	34029
ppptky391.asahi-net.or.jp	[01/Jul/1995:00:00:19	GET /images/launchpalms-small.gif HTTP/1.0	200	11473
205.189.154.54	[01/Jul/1995:00:00:24	GET /shuttle/countdown/ HTTP/1.0	200	3985
waters-gw.starway.net.au	[01/Jul/1995:00:00:25	GET /shuttle/missions/51-l/mission-51-l.html HTTP/1.0	200	6723
ppp-mia-30.shadow.net	[01/Jul/1995:00:00:27	GET / HTTP/1.0	200	7074

Fig. 2 Processed Data

Manuscript published on 30 March 2019.
*Correspondence Author(s)
Dharmendra Dangi, Department of Computer Applications, Maulana Azad National Institute of Technology, , Bhopal, India.
Amit Bhagat, Department of Computer Applications, Maulana Azad National Institute of Technology, , Bhopal, India.
Brijesh Bakariya, Department of Computer Science and Engineering, I.K. Gujral Punjab Technical University, Hoshiarpur, India.

Table 1 Predefined Timeslots

Timestamp (T)	Time Intervals
T1	00:01 04:00
T2	04:01 08:00
T3	08:01 12:00
T4	12:01 16:00
T5	16:01 20:00
T6	20:01 00:00

We introduce the proposed algorithm for pattern analysis from weblog. The flow of an approach is shown in above Figure 3.

Table 2 Transactional Record

Transactions (T)	URL 1	URL 2	URL 3	URL 4	URL 5	URL 6	URL 7	MF
T 1	0	1	0	1	0	0	0	1
T 2	2	0	0	6	2	0	5	6
T 3	1	2	0	6	1	5	0	6
T4	0	4	2	3	1	0	0	4
T5	1	2	1	0	1	0	2	2

According to follow of proposed approach as mentioned, we introduce an algorithm named Pattern Analysis based on Maximum Frequency of Weblog (PAMFW).

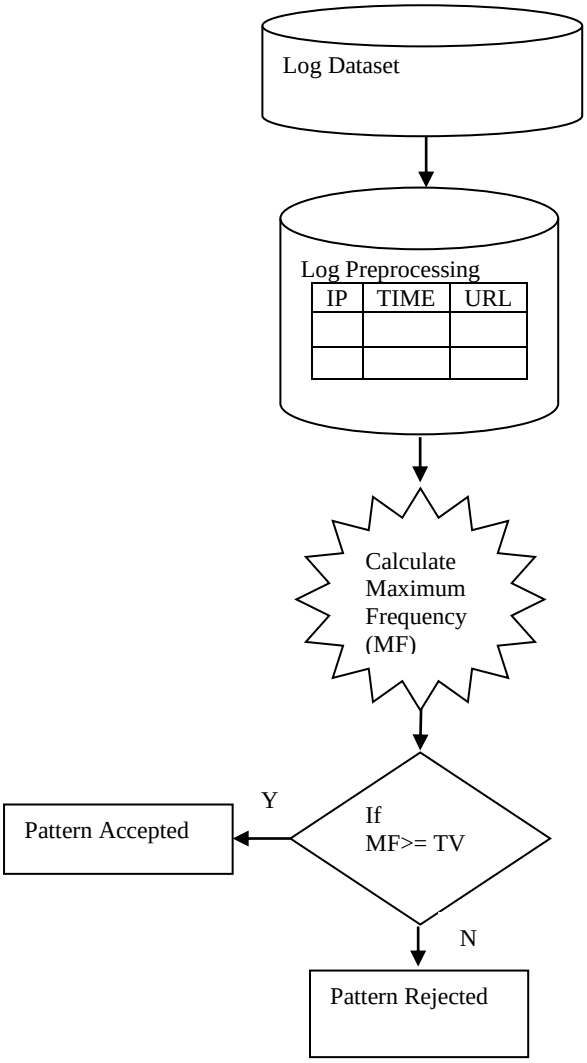

Fig. 3 Flow of Proposed Approach

Pattern Analysis based on Maximum Frequency of Weblog (PAMFW)

Terms using PAMFW:
Weblog W, Transactional Table TT.
Input:
Transaction Table TT, Threshold Value (TV)
Output:
All Strong Patterns based on MF

1: Start
2: Read TT, where TT ∈ W.
3: Calculate MF for every URL, where URL ∈ TT.
4: Calculate Max (MF) for TT (URL), where Max (MF) ⊆ TT ∈ W.
5: If MF (URL) ≥ TV i.e. favorable item (Fi), where Fi∈ TT.
6: Otherwise Unfavorable item (UFi) then Rejected UFi, where UFi ∈ TT.
7: Take Fi based on MF
8: Arrange according to MF
9: Stop

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

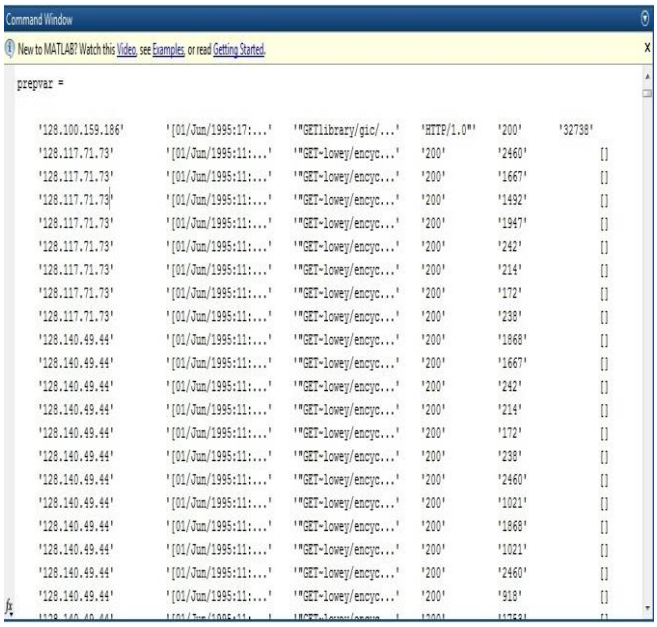

Fig. 4 Snapshot for Structured Weblog

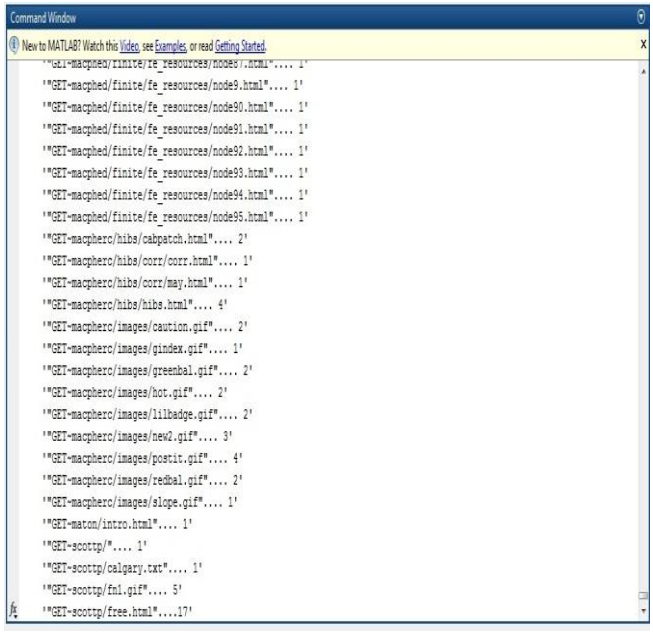

Fig. 5 Snapshot for Weblog Frequency

The experiment performed on 1000 we records and this dataset is downloaded from ITA Weblog repository. It is having 1000 web records, which contain various attributes of weblog. First of all we have taken one of the parameter "URL" or webpage from weblog data. First of all we have taken one of the parameter "URL" or webpage from weblog data and calculate its frequency. Here frequency means, how many time URL is repeated? First of all we have to find unique URL from weblog after that it have calculated the frequency of URL. We have implemented our proposed approach in MATLAB. In this experiment, the tests were conducted on a 1 thousand records. This dataset was extracted from the transactional database of a weblog dataset. The experiments based on threshold value. We have set the threshold value to 40% on 1000 web transactions. We compared our proposed algorithm with DefMe.

	URL	Frequency
2	"GET~scottp/publish.html"....	9
3	"GET~macphed/finite/fe_resources/node1.html"....	5
4	"GET~scottp/fn1.gif"....	5
5	"GET~friesend/tolkien/rootpage.html"....	4
6	"GET~lowey/webville/icons/help_32.gif"....	4
7	"GET~lowey/webville/icons/letter_32.gif"....	4
8	"GET~lowey/webville/icons/map_32.gif"....	4
9	"GET~macphed/finite/fe_resources/fe_resources.html"....	4
10	"GET~macpherc/hibs/hibs.html"....	4
11	"GET~macpherc/images/postit.gif"....	4
12	"GET~lowey/"....	3
13	"GET~lowey/encyclopedia/index.html"....	3
14	"GET~lowey/webville/icons/bus_32.gif"....	3
15	"GET~lowey/webville/icons/east_32.gif"....	3
16	"GET~lowey/webville/icons/north_32.gif"....	3
17	"GET~lowey/webville/icons/sound_32.gif"....	3
18	"GET~lowey/webville/icons/south_32.gif"....	3
19	"GET~lowey/webville/icons/taxi_32.gif"....	3
20	"GET~lowey/webville/map/citymap.gif"....	3
21	"GET~lowey/webville/map/index.html"....	3
22	"GET~macpherc/images/new2.gif"....	3
23	"GET~ladd/ostriches.html"....	2
24	"GET~lowey/encyclopedia/help.html"....	2
25	"GET~lowey/kevin.gif"....	2

MaxFrequency

Fig. 6 Structured Weblog Frequency

Figure 7 shown, the usage of memory on 40 threshold value. In this figure clearly demonstrate that DefMe. In this figure is clearly mentioned that PAMFW is quite better then DefMe according memory usage. Moreover according to space point of view our algorithm gives the fruitful result. In Figure 8, represent frequent itemset generation with different number of transactions. Here frequent itemset satisfies monotone and anti-monotone property. Here we set 40 as a minimum threshold value. In this figure clearly demonstrate that DefMe. In this figure is clearly mentioned that PAMFW is quite better then DefMe according frequent itemset generation.

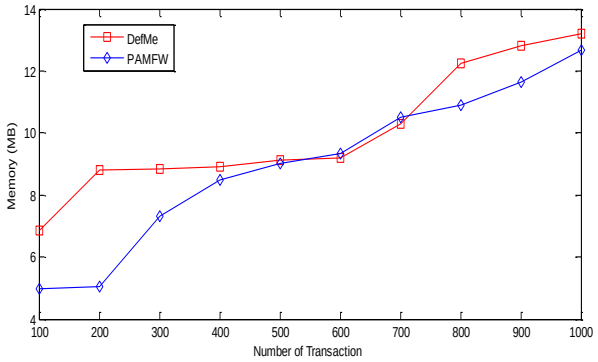

Fig. 7 Comparison with Transaction and Memory

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

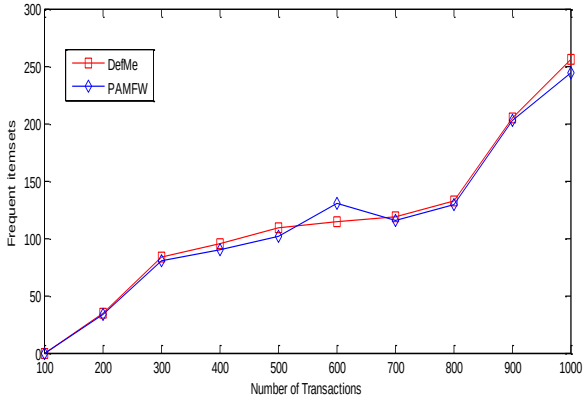

Fig. 8 Comparison with Transaction and Frequent Itemset Generation

In Figure 9, shows the execution time with different number of transactions. Here we fixed 40 as a minimum utility threshold value. In this figure clearly demonstrate that DefMe. In this figure is clearly mentioned that PAMFW is quite better then DefMe according frequent itemset generation.

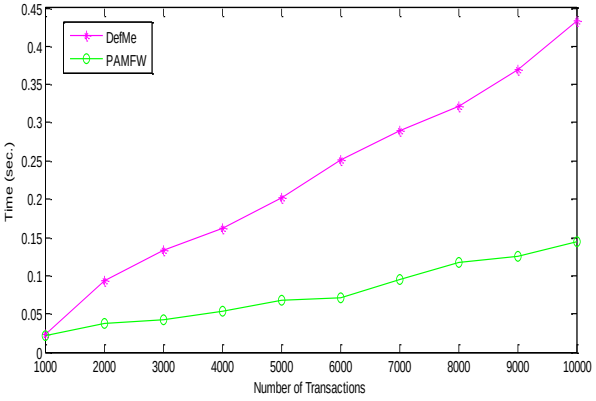

Fig. 9 Comparison with Transaction and Running Time

In this paper, we proposed Pattern Analysis based on Maximum Frequency of Weblog (PAMFW) algorithm and also proposed a framework for pattern analysis. This approach generates strong pattern from weblog. The weblog contain various number of parameter. We have compared our proposed approach PAMFW with the DefMe algorithm. We have analyzed our proposed approach if far better. The proposed approach is used for different areas such as Advertisement recommendation, E-Marketing, Business Intelligence, Education etc.

REFERENCES

1. D. Huang, Y. Koh, G. Dobbie, "Rare pattern mining on data streams. Data", *Warehousing and Knowledge Discovery Lecture Notes in Computer Science* 7448:303-314. doi: 10.1007/978-3-642-32584-7_25, 2012.
2. T. Papadopoulos, T. Stamati, P. Nopparuch, "Exploring the Determinants of Knowledge Sharing Via Employee Weblogs". *International Journal of Information Management* , Vol. 33(1): 133-146, 2013.
3. B. Bakariya, G.S. Thakur, "An Efficient Algorithm for Extracting Infrequent Itemsets". *The International Arab Journal of Information Technology (IAJIT)*, 16 (2), 2019.
4. R. Karim, C. F. Ahmed, B . Jeong, and H. Choi, "An Efficient Distributed Programming Model for Mining Useful Patterns in Big Datasets". *IETE Technical Review* 30:53-63, 2013.
5. B. Bakariya, G.S. Thakur, "Pattern Mining Approach for Social Network Service", *National Academy Science Letters, Springer*, 40(3):183–187, 2017.
6. Y. T. Wang, A. J. T. Lee, "Mining Web Navigation Patterns with a Path Traversal Graph". *Expert Systems with Applications, Elsevier*, 38:7112–7122, 2011.
7. B. Bakariya, G. S. Thakur, "Mining Rare Itemsets from Weblog". *National Academy Science Letters, Springer*, 39(5): 359–363, 2016.
8. http: //ita.ee.lbl.gov. Accessed 12 March 2013.
9. B. Bakariya, G. S. Thakur, "An Efficient Algorithm for Extracting High Utility Itemsets from Web Log Data". *The Institution of Electronics and Telecommunication Engineers (IETE) Technical Review*, 32(2):151-160, 2015.
10. A. Soulet, F. Rioult, "Efficiently Depth-First Minimal Pattern Mining", *Lecture Notes in Computer Science* 8443, 2014.

AUTHORS PROFILE

Dharmendra Dangi received Graduation degree from Jiwaji University, Gwalior in 2004, and Post Graduation Degree in SCSIT, DAVV, Indore from Devi Ahilya Vishwavidyalaya Indore M.P. in year 2013. He is UGC NET qualified in 2016 and six time GATE qulalifed from 2010 to 2016 He is Currently Pursuing PhD. Degree in the Department of Computer Applications, Maulana Azad National Institute of Technology Bhopal M.P. His Research interests include Big Data, Hadoop and Web.

Amit Bhagat has received his B.C.A and MCA degree in Computer Applications from Makhanlal Chaturvedi National University of Journalism, Madhya Pradesh in the year 2000 and 2003. He has done PhD from MANIT Bhopal in the year 2013. He is currently working as Assistant Professor in the Department of Mathematics and Computer Applications, Maulana Azad National Institute of Technology, Bhopal, and Madhya Pradesh, India. His research interests include Data Mining, Neural Networks, Sentiment Analysis, Web Mining and Big Data.

Brijesh Bakariya received Graduation degree from Barkatullah University Bhopal M.P. in 2005, and Post Graduation Degree in Computer Applications from Devi Ahilya Vishwavidyalaya Indore M.P. in year 2009. He received Ph.D. Degree in the Department of Computer Applications, Maulana Azad National Institute of Technology Bhopal M.P. in 2016. He is Assistant Professor in Department of Computer Science and Engineering, I.K. Gujral Punjab Technical University (IKGPTU) Jalandhar, Punjab. He has been teaching since 2009 and guiding M.Tech/ Ph.D students. In the mean time he published many research papers in SCI publications in the area of Data Mining, Image Processing, and Social Networking. He has attended various short term training programs, refresher course, workshops and seminars. He is a member of the IACSIT, APCBEES, APCBEES and UACEE.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)