Annals of DAAAM for 2011 & Proceedings of the 22nd International DAAAM Symposium, Volume 22, No. 1, ISSN 1726-9679
ISBN 978-3-901509-83-4, Editor B. Katalinic, Published by DAAAM International, Vienna, Austria, EU, 2011
Make Harmony between Technology and Nature, and Your Mind will Fly Free as a Bird
Annals & Proceedings of DAAAM International 2011

A METHOD FOR IMPROVING THE PREDICTION OF NEXT PAGE REQUEST OF A WEB USER

DINUCA, C[laudia] E[lena]; CIOBANU, D[umitru] & ISTRATE, M[ihai]

Abstract: *In this article we presented a way to improve the prediction of the next page request of a web user obtained with Page Rank algorithm. The used ideea is to apply Page Rank algorithm only on the subset of logs that contain the current page. To exemplify this method we use a set of web logs from the website of Nasa which is available online. We obtain an increase of visitation probability with 12.7%, from 19.8% to 32.5%*
Key words: *web logs, clickstream, page rank, prediction*

1. INTRODUCTION

A web site represents a set of interconnected web pages on the Web and is developed and maintained by a person or organization. While web sites constitute a medium for communication, publicity and commerce, Web Mining studies discover and analyze useful information from the Web (Nong, 2003).

Nowadays, there are many commercial and freeware software packages that provide basic statistics about web sites, including number of page views, hits, traffic patterns by day-of-week or hour-of-day, etc. These tools help ensure the correct operation of web sites (e.g., they may identify page not found errors) and can aid in identifying basic trends, such as traffic growth over time, or patterns such as differences between weekday and weekend traffic (Clark et al., 2006).

With growing pressure to make e-commerce sites more profitable, however, additional analyses are usually requested.

In this paper we present a method to help improve predictions of pages to be visited in order to create a recommendation system for web site users. We applied the Page Rank algorithm on NASA log file in order to get predictions for the next visited page. The method uses a table of probabilities of visited pages that are updated from time to time depending on the rate of visiting the website. This allows real-time calculation of visiting probability of the following pages considering the page where the user is at a specific time. The method can be used to create a recommendation system for web sites and pages preload to speed request response. The idea behind web caching is to maintain a highly efficient but small set of retrieved results in a cache, such that the system performance can be notably improved since later user requests can be directly answered from the cache. The recommendation should use the first three pages with the highest probability of visitation returned by the program.

2. DATA PREPROCESING

Log files are created by web servers and filled with information about user requests on a particular Web site. They may contain information about: domains, subdomains and host names; resources requested by the user, time of request, protocol used, errors returned by the server, the page size for successful requests.

Because a successful analysis is based on accurate information and quality data, preprocessing plays an important role. Preparation of data requires between 60 and 90% of the data analysis and contributes to the success rate of 75-90% to the entire process of extracting knowledge (Nong, 2003).

For each IP or DNS determine user sessions. The log files have entries like these:

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.

As can be noticed above, each record in the file contains an IP, date and time, protocol, page views, error code, number of bytes transferred. The steps needed for data preprocessing were presented in detail in (Dinuca, 2011). For sessions' identification in the first case was considered that a user can not be stationed on a page more than 30 minutes. This value is used in several previous studies, as can be seen in the work (Markov & Larose, 2007). The current study intends to add an improvement in sessions' identification by determining an average time of page visiting the sites for the visit duration determined by analysis of web site visit duration, data which can be found in the log files of the site. Thus, for each visited page, the visit duration is calculated, which is determined by the difference between two consecutive timestamps for the same user who is identified by IP. For records of pages with the highest timestamp among those visited by a user is assigned a predefined value of our choice to 20,000 seconds. We calculated the average visiting time for a page as the media of time spent for different users on that page and used this mean to better identify sessions. When calculating the average visiting time we don't take into consideration the pages with the time less than 2 seconds and largest than 20,000 seconds. For our analysis we selected only those log records that contained a web page, eliminating the required load images and other files adjacent to it, this information being considered not important for analysis. We kept only pages that have status code of class 200, a successfully loaded page. We have removed double pages from sessions and we just kept for review sessions with more than 1 viewed pages.

3. METHODS AND RESULTS PRESENTATION

We used to predict the next page visited the Page Rank algorithm. We consider the current session a session in progress and current page is the page that the user is at the time. To

improve the results we apply these methods only on sessions that contain the current page. From the all sessions we use about 85% for the calculation of the probability of visiting the page and on the rest sessions we check the accuracy of results.

For the first set of sessions we apply the Page Rank algorithm which provides us the ranks for pages from the websites. For each page we see on which pages can navigate and using the rank of pages we calculate the probability of visiting them by dividing each rank to the sum of ranks.

We implemented a program in Java using NetBeans IDE. It receives the log file in text format, write data for each session into a table, we code pages,we calculate the visiting time for each page, then calculate the average of each page visit, identify sessions, and apply Page Rank first on all chosen set of learning sessions and then only on the set of learning sessions that contain the current page obtaining the probabilities of visiting for first three more visited pages from where it can navigate from current page.

For the NASA data set we obtained after preprocessing 5138 sessions and we use 4486 for computing ranks and 652 for checking accuracy of the method. For each page we saved into a table the pages where it can navigate, pages with the highest probability of access obtained from the ranks of pages. In Fig.1. the table presents a part of the withhold visitation probabilities. So, from page 1 it goes with PR1 probability in page with CP_PR1 code, it goes with PR2 probability in page with CP_PR2 code and in page with CP_PR3 code with PR3 probability, PR means the probability obtained by applying Page Rank algorithm and CP stands from Page Code.

CP	PR1	CP_PR1	PR2	CP_PR2	PR3
1	0.5750204496079896	222	0.33796411569002527	294	0.0870154
2	0.9170015750273651	224	0.04500398937176131	74	0.0379944356
3	1.0	490	0.0	0	
4	0.8721891192825565	294	0.1278108807174435	23	
5	0.4046530030509877	129	0.2983883088683738	92	0.296958688
6	0.8654895144985767	189	0.13451048550142336	476	
7	0.8081620263579595	154	0.19183797364204042	377	
8	0.5006682327226563	264	0.29807482263923435	207	0.20125694
9	0.6642006433097454	62	0.3280382947950814	128	0.0077610618
10	1.0	358	0.0	0	
11	1.0	441	0.0	0	
12	0.6436991673143487	420	0.342106733546477	72	0.0141940991
13	0.5860213093835033	189	0.4139786906164967	490	
14	0.4046530030509877	129	0.2983883088683738	92	0.296958688

Fig. 1. Probabilities of visiting the pages obtained considering the whole data set of logs

Next, we use for each page in order to calculate ranks only sessions containing that page. Some of the ranks obtained can be seen in Fig. 2.

CP	PR1	CP_PR1	PR2	CP_PR2	PR3
1	0.1763024808025161	210	0.1663208180152588	366	0.1546916
2	0.3333333333333333	74	0.3333333333333333	14	0.333333
3	1.0	490	0.0	0	
4	0.6958272515961711	294	0.30417274840382896	23	
5	0.12780711914446785	422	0.1277780727012246	264	0.127370
6	0.5	476	0.5	189	
7	0.5829029381263467	377	0.4170970618736533	154	
8	0.413984482269055	301	0.21035765638876375	207	0.1152288
9	0.3074518076524626	467	0.26062405875122635	191	0.176352
10	1.0	358	0.0	0	
11	1.0	441	0.0	0	
12	0.5317316596105466	420	0.3024765097295488	444	0.1657918
13	0.6454048414151626	189	0.35459515858483737	490	
14	0.1979201730783784	92	0.15079585832790837	129	0.13587

Fig. 2. Probabilities of visiting the pages obtained considering only sessions containing the current page

The 652 sessions that were used to verify results have in total 3501 pairs of pages. From all of these, as can be seen in Tab. 1., 292 are verified by the highest ranking page, 186 page second page rank and 215 at the third rank. The last two columns from the table represent the sum of the first two columns and the sum of the first three columns.

pr1	pr2	pr3	pr 1+2	pr 1+2+3
292	186	215	478	693

Tab. 1. The number of correct predictions obtained when using the entire dataset

From the pages used to check data we obtained the data which can be seen in Tab. 2.

pr1	pr2	pr3	pr 1+2	pr 1+2+3
516	320	303	836	1139

Tab. 2. The number of correct predictions obtained by applying rank page only on sessions that contain the current page

Using the probability that the next visited page is among the three pages indicated from the program was 19.8% when we used all sessions and 32.5% when in the calculations we used only sessions containing the current page.

4. CONCLUSION

The method presented can be used online for prediction, recommendation and preload pages as the ranks are saved in tables and can be easily accessed in real time. The update of these tables is only required from time to time depending on the use of the site.

In the performed analysis the time was only used during the identification of sessions. For improved results in future research will take into account the order in which pages appear in the session and how long the current user staied on the visited pages before the current page.

5. REFERENCES

Clark L., Ting I., Kimble C., Wrigth P., Kudenko D. (2006), Combining Ethnographic and Clickstream Data to Identify Strategies Information Research 11(2), paper 249.

Cooley R., Mobasher B. and Srivastava. J. (1997), Web Mining: Information and Pattern Discovery on the World Wide Web. A survey paper. In *Proc. ICTAI-97*.

Database NASA Kennedy Space Center Log available online at http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html.

Dinuca C. E. (2011), The process of data preprocessing for Web Usage Data Mining through a complete example, Annals of the Ovidius University, Economic Sciences Series Volume XI, Issue 1 /2011.

Kohavi R., Parekh R. (2003), Ten supplementary analysis to improve e-commerce web sites, Proceedings of the Fifth WEBKDD workshop.

Liu B. (2006), Web Data Mining: Exploring Hyperlinks, Contents and Usage Data, Springer Berlin Heidelberg New York.

Markov Z., Larose D. T., (2007), DATA MINING THE WEB, Uncovering Patterns in Web Content, Structure and Usage, USA: John Wiley & Sons.

Nong, Y. (2003), The handbook of Data Mining, Lawrence Erlbaum Associates, Publishers Mahwah, New Jersey.

Robu, R.; Hora, C. & Stoicu - Tivadar, V. (2010). Improving the Classify User Interface in WEKA Explorer, Annals of DAAAM for 2010 & Proceedings of the 21st International DAAAM Symposium, 20-23rd October 2010, Zadar, Croatia, ISSN 1726-9679, ISBN 978-3-901509-73-5, Katalinic, B. (Ed.), pp. 0171-0172, Published by DAAAM International Vienna, Vienna.