On an Algorithm for Identifying Sessions from Web Logs

Claudia Elena Dinucă[1], Dumitru Ciobanu[2]

Abstract. The quality of decisions is based on the quality of processed data. So it is important that at the beginning of the data mining process to provide correct and quality data. The preprocessing data is a necessity for avoiding the failure of the data analysis. The idea that the data mining process can be done without human supervision has proved to be wrong. Even so, the humans are trying to automate as much as possible the process. From here are resulting many algorithms and techniques that are implemented using various programming language. In this work is presented an algorithm for identifying the sessions from a web logs file. It uses a value of 30 minutes to mark the end of a session and start another. We compute the average time for visiting the pages and using this we show that the presented algorithm produces errors in identifying sessions. We consider that the correct way to identify the session is to take into account the average time for visiting the pages.

Keywords: clickstream analysis; preprocessing data; sessions' identification.

JEL Classification: L86; C63; C88.

1. Introduction

World Wide Web or Web on short is the universal information space that can be accessed by companies, governments, universities, students, teachers, businessmen and some users. In this universal space trading and advertising activities are held. A Web site is a lot of interconnected web pages that are developed and maintained by a person or organization. Web mining and analyzing studies reveal useful information on the web. Web mining studies analyzes and reveals useful information from the Web (Cooley, Mobasher & Srivastava, 1997). Web mining is a term used for applying data mining techniques to Web access logs (Zaiane, 2000). Data mining is a non-trivial process of extracting previously unknown and potentially useful knowledge from large databases (Piatetsky-Shapiro, Fayyad, Smith & Uthurusamy, 1996).

[1] PhD Student, University of Craiova, Faculty of Economic and Business Administration, Romania, Address: A. I. Cuza, no. 13, Craiova, 200585, Romania, Tel: +4(251) 411317, Corresponding author: clauley4u@yahoo.com.

[2] PhD Student, University of Craiova, Faculty of Economic and Business Administration, Romania, Address: A. I. Cuza, no. 13, Craiova, 200585, Romania, Tel: +4(251) 411317, e-mail: ciobanubebedumitru@yahoo.com.

Web mining can be divided into three categories: Web content mining, Web structure mining and Web usage mining (Zaiane & Han , 1998). Web content mining is the process of extracting knowledge from documents and content description. Web structure mining is the process of obtaining knowledge from the organization of the Web and the links between Web pages.

Web usage mining analyzes information about website pages that were visited which are saved in the log files of Internet servers to discover the previously unknown and potentially interesting patterns useful in the future. Web usage mining is described as applying data mining techniques on Web access logs to optimize web site for users.

Click-stream means a sequence of Web pages viewed by a user; pages are displayed one by one on a row at a time. Analysis of clicks is the process of extracting knowledge from web logs. This analysis involves first the step of data preprocessing and then applying data mining techniques. Data preprocessing involves data extraction, cleaning and filtration followed by identification of their sessions.

2. Sessions Identification

Correct identification of sessions is an important step in preprocessing data from web logs. Some studies indicate a period of 30 minutes between pages viewed as sufficient to establish the end of a session and start another. However, this period may not be sufficient for certain types of websites, for example those which contains documents that the user reads. Also in this category may fall and commerce sites pages which are opinions about products. Should be taken into account that different people need time to cover the same amount of information, for example an elderly person can slowly follow the information presented on the website. Also in the case when a potential client who wants to better inform about a product may exceed this time and the analyst wrongly consider the session ended, longer time spent on the website in this case showing interest in the product and maybe the wish to purchase the product than to leave the website. More bad decisions in sessions' identification can significantly alter the results of applying data mining techniques. In an attempt to reduce errors in session identification, an improved algorithm is proposed to amend the classic algorithm. More bad decisions identification sessions can significantly alter the results of applying data mining techniques.

In an attempt to reduce errors in sessions identification we propose to amend the sessions identification algorithm.

Model description.

We consider IP the set of IP adresses of users = {IP1, IP2, ..., IPN}. PIPk is the set of user pages that were visited by the user identified through IP_k IP, PIPk = {PIPk1, PIPk2, ...} and TS_PIPki is the timestamp of PIPki page. We denote by ID_PIPki the sessions identifications numbers assigned to PIPki page and we note ID the set of all these identifications numbers.

The pseudo-code Algorithm

For each IP IP_k repeat

If | PIP_k |=1 and ID_PIP_{k1}=max(ID)+1;

Then ID_PIP_{k1}=max(ID)+1;

I=1;

While (I<| PIP_k |) repeat

I=I+1;

If TS_PIP_{ki}- TS_PIP_{ki-1}<1800 then ID_PIP_{ki}= ID_PIP_{ki-1};

Else ID_PIP_{ki}= ID_PIP_{ki-1}+1;

In the logs table from the database we create a column to keep the time that user spent on the page regardless of session. We select the pages for each IP ordered by timestamp of the IP we make the difference between timestams of consecutive pages. For the last page we attribute a great value for example 20,000 seconds. Now we can calculate in various ways an average time that user spent on a web page. We set a maximum time limit of 2 hours time for the visit allocated to a page visit and a minimum of 2 seconds. We eliminate records that are off limit and calculate the average time spent by an user on a page. Based on this average time we will decide if the page is part of the old session or it is the first page in a new one.

If the average time spent by users on that page is close to 30 minutes it is clear that the algorithm presented above will produce errors in identifying sessions.

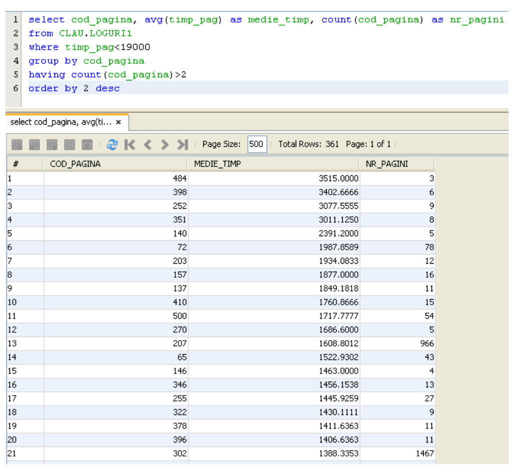
3. Case Study

We used the logs database that can be free downloaded from NASA website by clicking on http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html. For the implementation we used Java programming language. We used the version Java jdk.1.6.0. It is used NetBeans IDE, version 6.9.1. We calculated how long an user could stay on a page. For this we proceeded as follows. First we selected all the distinct IDs. For each ID we selected the identifications codes for each visited pages and the timestamp. When we have found some pages accessed only one time we attribute a default value of 20000 seconds. When we have more viewed pages, we calculate the time as the difference between two consecutive timestamps and for the last page we would set the default value of 20000. After data preprocessing phase there have been obtained 47 583 for 508 separate pages and 12 805 distinct IDs.

From these 508 pages, 118 pages were visited only once or twice.

To calculate the average time spent on a page we have eliminated times greater than 19000 seconds and we grouped by the codes of pages.

In Fig. 1. we display the pages in descending order of average time spent on those pages by users. The fields displayed in Fig. 1. are cod page (COD_PAGINA), average time (MEDIE_TIMP) and the number of visits (NR_PAGINI) for the page identified by cod page. Thus for the 14 pages that the average time of visiting is more than 1500 seconds, the probability to assume wrongly a session end is very high. We will look more closely at page 207 that has the most visits and the average visitation time is 1608.80 seconds.



```
select cod_pagina, avg(timp_pag) as medie_timp, count(cod_pagina) as nr_pagini
from CLAU.LOGURI1
where timp_pag<19000
group by cod_pagina
having count(cod_pagina)>2
order by 2 desc
```

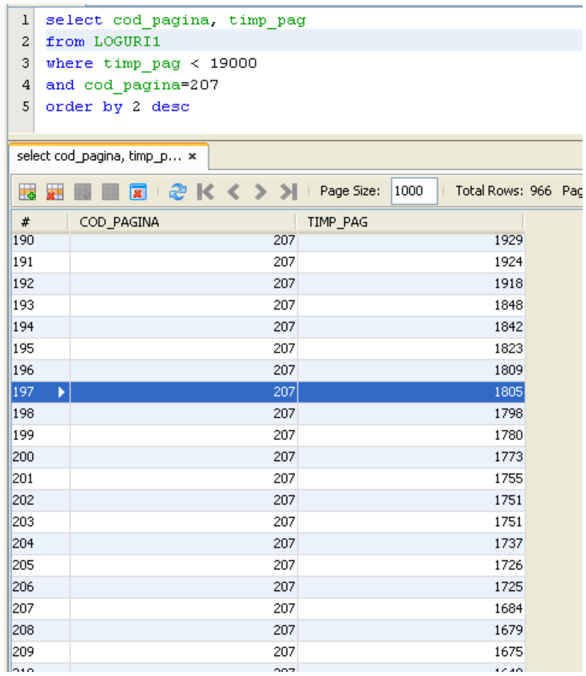
select cod_pagina, avg(ti... ×

Page Size: 500 Total Rows: 361 Page: 1 of 1

#	COD_PAGINA	MEDIE_TIMP	NR_PAGINI
1	484	3515.0000	3
2	398	3402.6666	6
3	252	3077.5555	9
4	351	3011.1250	8
5	140	2391.2000	5
6	72	1987.8589	78
7	203	1934.0833	12
8	157	1877.0000	16
9	137	1849.1818	11
10	410	1760.8666	15
11	500	1717.7777	54
12	270	1686.6000	5
13	207	1608.8012	966
14	65	1522.9302	43
15	146	1463.0000	4
16	346	1456.1538	13
17	255	1445.9259	27
18	322	1430.1111	9
19	378	1411.6363	11
20	396	1406.6363	11
21	302	1388.3353	1467

Fig. 1.

From the 966 visits of page 207, 197 visits have visited time greater than 1800 seconds (Fig.2.) and can lead to errors in the sessions' identification.



```
select cod_pagina, timp_pag
from LOGURI1
where timp_pag < 19000
and cod_pagina=207
order by 2 desc
```


#	COD_PAGINA	TIMP_PAG
190	207	1929
191	207	1924
192	207	1918
193	207	1848
194	207	1842
195	207	1823
196	207	1809
197	207	1805
198	207	1798
199	207	1780
200	207	1773
201	207	1755
202	207	1751
203	207	1751
204	207	1737
205	207	1726
206	207	1725
207	207	1684
208	207	1679
209	207	1675

Fig. 2.

Last observation justifies the proposal to replace the value of 1800 seconds (30 minutes) from the session' identification algorithm with another value that depends on the average time.

4. Conclusions

For a successful analysis of click-stream it requires the use as accurate data from web clicks. Sessions identification is an important step in data preprocessing whose poor performance may negatively influence the results. We note that the determination of the main visiting time of web pages from the websites requires, depending on the size of log files used for a certain period of time that is unprofitable to determine in real time. But calculating the mean can be offline and may be updated, depending on the level of accessing the website, daily, weekly or even less. Using a calculated time depending on average time for sessions' identification increases the accuracy of data used in the knowledge extraction process. It remains an open problem on a different calculation method of the time based on the mean time to have maximum effect when used to identify sessions.

5. References

Brendt B., Spiliopoulou M. (2000). Analysis of Navigation Behaviour in Web Sites Integrating Multiple Information Systems. VLDB, 9(1), 56-75.

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., Srivastava J. (1997). Web mining: Information and Pattern Discovery on the World Wide Web. A survey paper. In: Proc. ICTAI-97.

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

Hay B., Geert W., Koen V.(2005). *Discovering interesting navigations on a web site using SAM[I]*, Springer-Verlag Berlin.

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

Li T. R., Xu Y., Ruan D., Pan W. M. (2005). *Sequential Pattern Mining*. Springer-Verlag Berlin.

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

Mobasher B., Cooley R., Srivastava J. (1999), Creating Adaptive Web Sites trough usage based clustering of URLs, IEEE knowledge & Data Engg work shop (KDEX'99).

Piatetsky-Shapiro g., Fayyad U., Smith P., Uthurusamy R. (1996), Advances in Knowledge Discovery and Data Mining., AAAI/MIT Press.

Srivastava J., Cooley R., Deshpande M., Tan P.-N. (2000), Web usage mining: discovery and applications of usage patterns from web data, SIGKDD Explorations, 1(2), 12-23.

Zaiane O. (2000) Conference Tutorial Notes: Web Mining: Concepts, Practices and Research. In: Proc. SDBD-2000, 410-474.

Zaiane O., Han J. (1998), WebML: Querying the World Wide Web for resources and knowledge. In: Workshop on Web Information and Data Management WIDM98, Bethesda, 9-12.