Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 3 (2011) 677–683

Procedia Computer Science

www.elsevier.com/locate/procedia

WCIT-2010

Probability modeling of accesses to the web parts of portal

Michal Munk [a], Marta Vrábelová [b], Jozef Kapusta [a] *

[a] Constantine the Philosopher University in Nitra, Department of Informatics, Tr. A. Hlinku 1, 949 74 Nitra, Slovakia
[b] Constantine the Philosopher University in Nitra, Department of Mathematics, Tr. A. Hlinku 1, 949 74 Nitra, Slovakia

Abstract

The analysis of behavior of portal visitors is one of the most important parts of web portal optimization. The results of the analysis are important for the further correction and improvement of web part organization. The aim of the paper is modeling of probabilities` accesses to the categories of web parts of portal. We deal with the access probabilities to the individual categories of faculty portal content depending on the day's hour and the week's day. The probabilities are estimated using multinomial logit model for employees and students separately. In logit models, in case of students and employees, the week's days present statistically significant signs, representing dummy variables (MON, TUE, …) in the model. On the other hand, day's hours representing with variables HOUR_DAY and their square HOUR_DAY_Q, are shown as statistically significant signs only in the case of students. These results correspond with the computing probabilities wherein the probabilities of access to web parts of the portal are more stable in the case of employees than of students during the day. The analysis provided us several interesting and surprising results. For instance, from the analysis, results follow that the part *study* is the most visited part by students in the evening and night hours. The analysis results confirmed general trends, for example the part *announcements* is the most visited part in morning's hours, at the beginning of the week especially. All of the analysis results will help us to further optimize our web portal. This is especially point in level of portal adaptivity on the basis user and access hour on portal.

Keywords: web log mining, user behavior, web parts, probability modeling, multinomial logit model ;

1. Introduction

The one of the first of the web mining task is an analysis of behavior of portal visitors. On the basis of the found knowledge, it is possible to correct and improve web page organization. The aim of the paper is the probability modeling of accesses to the categories of the web parts of portal. For this purpose, data about accesses to the categories of the web parts of the Faculty of Natural Sciences of Constantine the Philosopher University (FNS CPU) portal were collected. We deal with the access probabilities to the individual categories of faculty portal content depending on the day's hour and the week's day. The probabilities are estimated using multinomial logit model [1], [2] for employees and students separately.

* Jozef Kapusta. Tel.: +421-37-6408-675 ; fax: +421-37-6408-556 .
E-mail address: jkapusta@ukf.sk .

1877-0509 © 2010 Published by Elsevier Ltd.
doi:10.1016/j.procs.2010.12.113

2. Data Source

The automatically saved data in log file are our data source. For this reason, this area is also often known as web log mining. In this data we follow the sequences in visiting each page of user, who is identified by defined conditions, such as the IP address. Log file, in its standard structure called Common Log File [3] marks each transaction performed by the browser in accessing the web. Each row presents notice about IP address, the time and the date of the visit, the approaching object and referring object. If we use its extended image, we may also record the data about the version of the browser, User-Agent.

3. Data Preprocessing

The log file of the web server is the source of anonymous data about the user. On the other hand, these anonymous data present problem by unambiguous identification of the particular user. For this purpose the following adjustments (corrections) are made:

- Data cleaning from the crawlers of search services accessed to the portal [4].
- Identification of visitors, on the basis of the various internet browsers [5].
- Identification of sessions, where the session may be defined as a sequence of the steps, that lead to completing the concrete task [6] or as a sequence of the steps, that lead to meeting the concrete target [7]. The simplest method is if we consider the series of clicks in a defined period of time, for example 15 minutes [8].
- The reconstruction of activities of a web visitor. Taucher and Greenberg [9] proved that more than 50% of accesses to web are via backward path. Here comes the problem with the cache of the browser. By the backward path, a query for web server is not running, thus there does not exist a record in the log file. The solution to this problem is path filling. With path filling we add these missing rows into the log file [5].

All of these techniques depend on the first step of data collection and data cleaning [10].

For the needs of research, it was required to classify the visitors into separate groups. By each portal visitor we can find out his/her IP address. Within the institute, we can define according to IP address, whether it belongs to employee or student. The disadvantage is, we could only define the employee or student from accesses from the university network, meaning from the employees` PCs at the workplace and students` PCs in PC rooms and dormitories.

4. Variables

The investigated categorical dependent variable is a variable CATEGORY with categories: *introduction, study, announcements, about faculty, information for, information on, regulations - public notices, documents, science - research, others, retrieving, events, conference*s, and the categories were chosen on the basis of their competency to context. The variable HOUR_DAY with values 0-23 is non-dependent variable. We use week days MON, TUE, WED, THU, FRI, SAT as dummy variables. The data describes individual accesses on the FNS CPU portal in two weeks from the middle of the summer semester.

5. Description of the Model

Denote by π_{ij} the probability of choosing the web portal category j by a visitor in the hour i, where $j = 1, 2, ..., J$. Since $\sum_{j=1}^{J} \pi_{ij} = 1$, there are J - 1 parameters.

Let Y_{ij} be the number of accesses in the hour i to category j with observed value y_{ij}. Then $n_i = \sum_j y_{ij}$ is the number of accesses in the hour i. The probability distribution of Y_{ij}, in the case n_i is given, is multinomial,

$$P\left[Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{iJ} = y_{iJ},\right] = \frac{n_i!}{y_{i1}!\, y_{i2}! \ldots y_{iJ}!} \pi_{i1}^{y_{i1}} \pi_{i2}^{y_{i2}} \ldots \pi_{iJ}^{y_{iJ}} . \qquad (1)$$

The probabilities π_{ij} of choosing category j with respect to the hour i we get from multinomial logits which we will model. The multinomial logits are logarithms

$$\ln\frac{\pi_{ij}}{\pi_{iJ}},\ j=1,2,\ldots,J-1,\ i\in\{0,1,\ldots,23\},$$

where π_{iJ} is the probability of last (reference) category. We assume the following model

$$\eta_{ij}=\ln\frac{\pi_{ij}}{\pi_{iJ}}=\alpha_j+\mathbf{x}_i^{'}\beta_j, \tag{2}$$

where $\mathbf{x}_i^{'}$ is a line vector, α_j is a constant and β_j is a vector of regression coefficients, for $j=1,2,\ldots,J-1$. There are J - 1 equations which contrast each of categories $1,2,\ldots,J-1$ with the category J. The probabilities π_{ij} we obtain from formulas

$$\pi_{iJ}=\frac{1}{1+\sum_{k=1}^{J-1}e^{\eta_{ik}}},\quad \pi_{ij}=e^{\eta_{ij}}\pi_{iJ},\ j=1,2,\ldots,J-1. \tag{3}$$

Maximum likelihood estimation of the parameters of the model (2) proceeds by maximization of the multinomial likelihood (1) with the probabilities π_{ij} viewed as functions of α_j and β_j. The estimation of the parameters can be done for individual data in a statistical system [2]. Then $n_i=\sum_j y_{ij}=1$ for $i=1,2,\ldots,n$, where n is the number of all accesses.

6. Determination of the Model and Parameter Estimations

We have found out with respect to contingence tables that the significant numbers (more then 10) of accesses for variables CATEGORY, HOUR_DAY and DAY_WEEK are

- for employees in the working days MON – FRI, in the hours 8-15, to the categories *introduction, study, announcements, about faculty,*
- for students in days MON – SAT, in hours 7–22, to the categories *introduction, study, announcements, about faculty.*

We have excluded non-significant accesses. The numbers of accesses to those categories are in table 1.

Table 1. The numbers of accesses to categories

Category	Numbers of accesses	
	employees	students
Introduction	758	2245
Study	818	972
Announcements	111	164
About faculty	287	119
Sum	1974	3500

We have used the following models (the variable HOUR_DAY is denoted by *t*).
For employees:

$$\eta_{ij}=\alpha_j+\beta_{1j}t_i+\beta_{2j}t_i^2+\gamma_{1j}MON_i+\gamma_{2j}TUE_i+\gamma_{3j}WED_i+\gamma_{4j}THU_i \tag{4}$$

For students:

$$\eta_{ij} = \alpha_j + \beta_{1j} t_i + \beta_{2j} t_i^2 + \gamma_{1j} MON_i + \gamma_{2j} TUE_i + \gamma_{3j} WED_i + \gamma_{4j} THU_i + \gamma_{5j} FRI_i \qquad (5)$$

The parameters of the models were estimated for the individual data in the *STATISTICA* software. They are in the table 2. The significance of parameters was tested using *Wald test*; significant parameters are colored.

Table 2. The parameters estimations

	Category	employees	students	Category	employees	students	Category	employees	students
Intercept	Introduction	6,840	1,348	Study	6,137	-2,849	Announcements	2,691	1,515
HOUR_DAY	Introduction	-0,978	0,342	Study	-0,926	0,384	Announcements	-0,821	-0,383
HOUR_DAY_Q	Introduction	0,041	-0,016	Study	0,039	-0,009	Announcements	0,030	0,012
MON	Introduction	0,258	-0,634	Study	0,424	0,955	Announcements	2,266	1,651
TUE	Introduction	0,208	1,103	Study	0,949	3,028	Announcements	2,985	3,049
WED	Introduction	-0,642	0,114	Study	-0,154	1,548	Announcements	0,550	1,591
THU	Introduction	-0,916	-0,499	Study	-0,011	1,221	Announcements	-0,611	0,143
FRI	Introduction		0,687	Study		1,208	Announcements		-1,242

By the table 2 the logits for the category introduction are dependent on the access hour and on the square of access hour. The values of logits for employees are significantly influenced by Monday and Thursday. Monday and Tuesday affect the values of logits for students. Interpretations of parameters for other categories are similar.

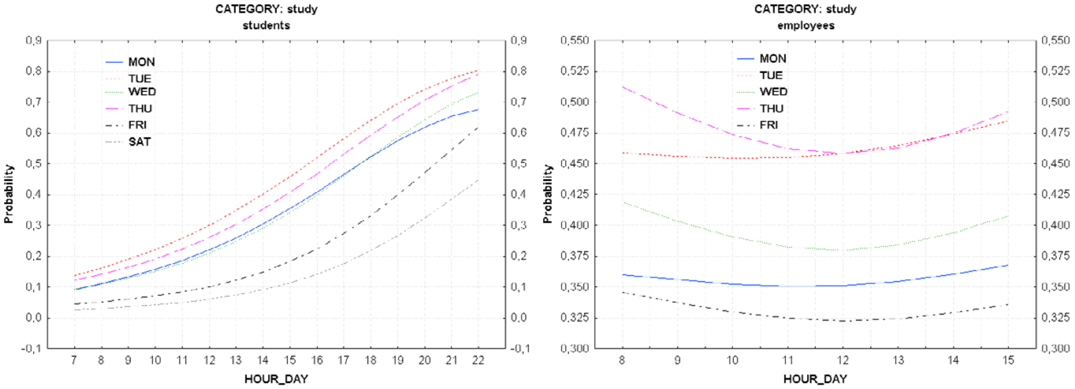

Fig. 1. Probabilities of accesses to the category study: (a) for students; (b) for employees

By the estimated parameters the estimates of logits and the probabilities of choosing individual categories in a any given day's hour can be enumerated. Then we can enumerate the theoretical numbers of accesses to individual categories. We give some illustrations. Graphs of the probabilities of choice of category *study* by students are in the figure 1(a) and the figure 1(b) obtains these probabilities for employees. They are very different for students and for employees. The probabilities for students increase from the morning to the evening, probabilities for employees are more stable. This corresponds with the *Test of all effects* and the *Likelihood type 3 test* results where all effects are significant for the students, but the day's hour and the day's hour squared are not significant in the case of the employees. The student chooses the category *study* with the greatest probability on Tuesday and this probability increases from 0.138 (7.00) to 0.803 (22.00). The employee chooses the category *study* with the greatest probability on Thursday, the probability is equaled to 0.512 (8.00), it decreases to 0.458 (12.00) and then it increases to 0.493 (15.00). We get interesting results for the probabilities of choice of different categories in one day by students. The probabilities of access to the categories on Wednesday for students are drawn in the figure 2(a). The students click on *introduction* in the morning with probability 0.773 (7.00) and this probability falls to 0.156 in the evening

(22.00). The probability of access to *study* is equaled to 0.093 in the morning and it grows to 0.73 in the evening. Differences between empirical and theoretical frequencies of accesses to the categories on Wednesday for students are drawn in figure 2(b). There is a problem at 12 o'clock, the probability of access to *introduction* is overestimated and the probability of access to study is underestimated. The value of difference at 12 o'clock for *introduction* is -19.89, it is an outlier since the standard deviation of these differences equals to 6.89, the average is 0 (-19.89 < 0 – 2*6.89). The value of difference at 12 o'clock for *study* is 20.131, the standard deviation equals to 7.116, the average is 0 (20.131 > 0 + 2*7.116) and so this difference is an outlier too.

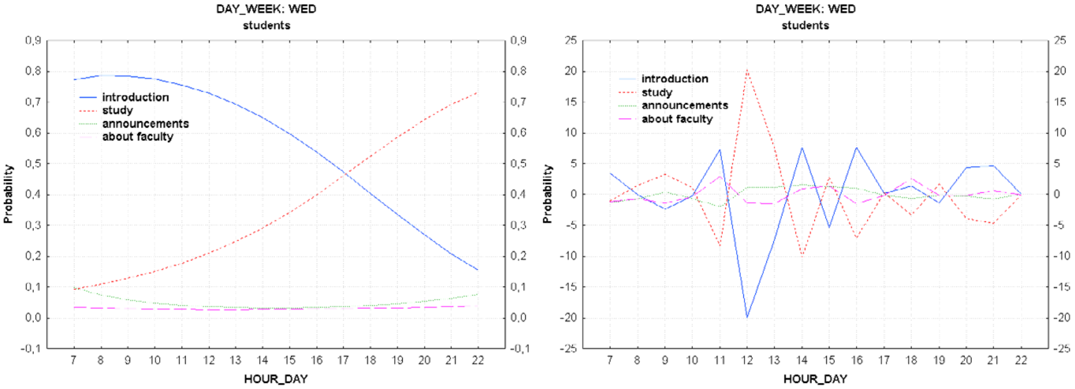

Fig. 2. Accesses to the categories on Wednesday for students: (a) probabilities; (b) differences of empirical and theoretical frequencies

7. Conclusions

In the article, we wanted to show the modeling possibilities of the distribution of categorical variable with the support of a multinomial logit model. As an example of the categorical variable we chose variable CATEGORY, where the levels are the categories of the web parts of the FNS CPU portal. We created the model for the employees, students, while we used week`s day and day`s hour as a predictive variables.

On the basis *Test of all effects* and the *Likelihood type 3 test* in created logit models, results present the week`s days statistically significant signs, which are represented with dummy variables in the models (MON, TUE, ...). Vice versa, week`s hours represented by variables HOUR_DAY and their square HOUR_DAY_Q, were shown as statistically significant signs only in the case of students. These results correspond with the counted probabilities, where the probabilities of access to the web parts of portal are more stable by employees than by students during the day. In models, the quite high correct prediction of classification of cases into the category *study* (71%) was reached in case of employees. And up to 93% correct cases` classification into the category *introduction* in the case of students.

From the results of the analysis follow that the part *study* is the most visited part by the group students in the evening and night hours. It was surprising, because this web part includes information connected with the study, study administration etc.. We assumed that this exact part would be the most visiting part during the day, when the students are at school. The general student trend as a web visitor seems to be that the students familiarize themselves with important announcements, which are located on the home page and they attend to study-related activities when they have more time, meaning in the evening and at night. The analysis results confirmed also the general trend that the *announcements* is the most visited area during the morning's hours, at the beginning of the week especially. We have adapted the politics of filling the web and releasing the announcements, where we try to release all of the important announcements by Monday morning latest.

In spite of the interesting analysis results, which were proved by analysis, some of the results do not have to be so significant. For instance, the visiting trend of the home page during the day certainly distorted by the fact that the faculty`s page is installed as the homepage in most of the computers in the classrooms. Each connection to a web browser it is automatically recorded into the log file, no matter if the user wants to visit that page currently. All acquired knowledge may be used as a source for the further analysis and optimization of the portal. The thought optimization and corrections of the portal will be realized in these following levels:

- Personalized access on the basis of user - portal of faculty, e.g. more similar portals, has behind long time running during which enriches constantly new information but with the increasing capacity it loses the transparency of the portal. Several years the authors of the faculty`s portal have endeavored to not take a long time to find necessary information for the users. For this reason the portal was restructured two years ago. Therefore information currently integrated into the parts like *event*s, *announcements* and *conference*s are now displayed on the home page. Similar, the main menu was restructured. We realize that these parts are always defined for concrete group of visitors, for example in the web part *announcements*; there is mainly information for students and etc. The endeavor of portal personalization is considered on two levels:
  - Authorized access – in spite of the compass of authorized access with the assistance of data access assignment, it exists, but its practical usage is reduced to employees with the form of extended part of main menu about several items. The endeavor is to personalize the portal for the employees (this group is possible to further divide) and students after their log in and other visitors (non- authorized user). The problem of authorization is its own log in, that mean "to force" the user to authorize and "not to burden" needlessly the users simultaneously.
  - Non- authorized access – we know how to find out the IP address of each portal visitor by using available existing technologies. We can define according to IP address and determine if it belongs to an employee or a student within the institution. From the point of view of the problem of authorized access, it is possible to personalize the portal according to the user`s IP address. Downside to this is that this method of access incorrectly identifies employees and students who visit the portal from home, meaning that they do not access the portal from the university network.

  For completeness, we add, that personalization itself will be implemented by adapting the following areas of the homepage: *events, announcements, conferences* and the main menu. We plan to use the technique of *sorting links* (or part of page) and *hiding links* (or part of page) as the techniques for personalization.
- Web parts correction (possible infunctionality) – we can find out the attendance of the separate page dependence on the hour of the concrete day with the help of the created model. This knowledge may be used for planning corrections, portal or its web parts layoff. Thus the portal is used as an interface (entry point) for access to the other institution`s systems and on the basis of that acquired knowledge we can deliver this information to the administrators.

References

1. J. Anděl. Základy matematické statistiky. MATFYZPRESS, Praha 2007. ISBN 80-7378-001-1.
2. G. Rodríguez. Generalized Linear Models. 2007. [cit. 2010-05-11]. Available from: http://data.princeton.edu/wws509/stata
3. W3C. Configuration File of W3C httpd [online]. 1995. [cit. 2009-03-10]. Available from: http://www.w3.org/Daemon/User/Config/Logging.html
4. A. G. Lourenço and O. O. Belo. Catching web crawlers in the act. Proceedings of the 6th international Conference on Web Engineering, 2006, ICWE '06, Vol. 263, ACM, New York, NY, pp. 265-272. ISBN 1-59593-352-2.
5. R . Cooley and B. Mobasher, J. Srivastava. Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information System, 1999, Springer-Verlag, Vol. 1, ISSN 0219-1377.
6. M. Spiliopoulou and L.C. Faulstich. WUM: A Tool for Web Utilization Analysis. Extended version of Proc. EDBT Workshop WebDB'98, 1999, Springer Verlag, pp 184–203.
7. M. Chen and J. S. Park, P. S. Yu. Data mining for path traversal patterns in a web environment. ICDCS, 1996, pp. 385–392. ISBN 0-8186-7398-2.
8. B. Berendt and M. Spiliopoulou. Analysis of navigation behaviour in web sites integrating multiple information systems. The VLDB Journal, 2000, Vol. 9, No. 1, pp. 56-75. ISSN 1066-8888.
9. L. Taucher and S. Greenberg. Revisitation patterns in world wide web navigation. Proc. of Int. Conf. CHI'97, 1997, Atlanta.
10. M. Munk and J. Kapusta, P. Švec. Data Preprocessing Evaluation for Web Log Mining: Reconstruction of

Activities of a Web Visitor. Procedia Computer Science, 2010, Vol. 1, No 1., pp. 2267-2274. ISSN 1877-0509.