A Discrete Choice Model for Web Site Work Results

Pavlyuk, Dmitry

Transport and Telecommunication Institute

24 August 2008

A Discrete Choice Model for Web Site Work Results

Pavlyuk Dmitry
Candidate of Economic Science, Transport and Telecommunication Institute
Lomonosova 1, Riga, LV-1019, Latvia
Phone: (+371)29958338. Email: dmitry@magicwebsolutions.co.uk

ABSTRACT

Currently a corporate web site is not considered as a necessary business attribute, but as a marketing tool which should yield results. In this study we consider a web site as an instrument for attraction of new partners (customers, suppliers). Web site outputs are a number of visitors interested in contact information (reached the contact info page) and a number of visitors who sent a request via a special form on the web site.

We build a sequential discrete choice model for web site outputs. Explanatory variables set includes a number of visited pages, seconds spent on the web site, and dummy variables for specific pages visited (a page with prices information, a portfolio page). Also we investigate an influence of search engines (Google, Yahoo, MSN), which refer a visitor to a corporate web site and keywords used for pay-per-click advertising campaigns.

We estimate model parameters on the base of a small UK-based web development company's web site statistical data and discover strong dependencies, which allow improving web site organisation and its search engine positioning.

Keywords: sequential discrete choice model, corporate web site

Introduction

Times when a corporate web site was considered as a necessary business only are gone. Today a web site is an important business and marketing tool, and it should yield results.

The results of web site functioning strictly depend on its goals. There are many different types of web sites, and each one has own primary goals. For example, a community web site should attract people and force them to become regular visitors, a product web site should provide full information about the product and its support, and so on. A corporate web site also can have different goals – providing information about a company and its services, attracting new partners, being a simple business card.

In this study we consider a very specific type of web sites – a corporate web site, oriented to receiving orders from new clients. This type of web site is very popular for offshore services, because it is the main (and often the only) way to receive new orders in this case. That's why such companies pay huge attention for their web sites and spend significant resources to get new visitors. The efficient, excellent structure of the web site is very important in this situation.

There are many different tools used to analyse web site logs (data storages with full information about site visits) and discover visitors' behaviour. Such tools allow to find popular visitors paths on the web site and exit points (web pages after visiting which visitors leave the site). The most of tools provide excellent descriptive statistics, but further statistical and econometric analysis is also definitely required.

There are many factors affecting a visitor of the web site. We suggest separating them into two groups:

- "visible" (in log files) factors;
- "invisible" factors.

"Visible" factors can be presented by a set of indicators mined from the log file. How many pages were visited? How much time is spent on the site? Which pages were visited? Where does a visitor come from? All these questions can be answered on the base of log files and used for further analysis.

"Invisible" factors affect visitor's behaviour, but can't be discovered from the technical statistics. For example, a visitor can leave an excellent web site, which is very important for him, just because he doesn't like its colours, or his internet connection is broken, or even due to an exciting Olympics event at that time.

Another difficulty is related with results of a web site visit. There are also some result types "visible" in log files – he visited a specific page (usually a page with contact information), sent an enquiry via a form on the site or clicked on an email link. But some positive results can't be found in the log file. For example, a visitor can note a phone number or an address on the web site and call or visit the office. Such results are technically "invisible". We use only "visible" factors and results in our econometric model.

Visitor's behaviour on the web site is a set of decisions. We mark out next critical (from business point of view) decisions:

- – to leave the site just from the first page or to continue browsing;
- – to view contact information or it is unnecessary;
- – to send an enquiry or to leave the site.

We built a sequential discrete choice model [1, 2], which allows to discover factors which significantly affect the decision on each stage and predict visitors' behaviour. The model uses log file information only and can extent standard web analytics tools.

Methodology and Data

We use a discrete choice model for web site visitor decisions. The mathematical formalization of the model can be presented as:

$$P(y = 1 | x = X) = F(X^T \beta),$$

where

y – a discrete (binary in our case) variable, which equals 1 if a visitor accepts an alternative and 0 if he declines it;

x – a set of explanatory indicators;

β – a vector of unknown coefficients to estimate;

F – a function, transforming R into [0, 1]. Usually F is a normal cumulative distribution function (the model is called "probit" in this case) or a logit function (the model is called "logit" respectively).

Discrete choice models are a very popular for modelling customers' decisions, the full description can be found at [1].

The decision is multistage in our case (see Diagram 1). At each stage a visitor chooses one of two alternatives – exit or continue web site browsing (there are numbers of visitors who chose an alternative in square brackets on the diagram).

A visitor comes to the web site from a search engine or by entering the site URL into the browser address line directly. The visit can be accidental (the visitor doesn't interested in web site content) or the visitor doesn't like the web site from the first glance. In this

situation we have no information about visitor's behaviour yet, and that's why we decided to exclude this stage from considering.

At the second stage the visitor decides (or doesn't decide) to visit a page with contact information. We assume that the majority of contact info visits are purposive. The decision to visit a page is the first level decision in our model.

The last, third, decision is made when a visitor fills a form. A form can be filled with info about a new potential project (an enquiry), with a message from an existed customer, or with a spam message. As we interested in new clients only, we excluded browsing sessions with latter variants (an existed client and a spam) from the sample.

```
          ┌─────────────────────────────────┐
          │  Visited the web site (one page) │
          │              [4742]              │
          └─────────────────────────────────┘
                          │
             ┌────────────┴────────────┐
             ▼                         ▼
      ┌────────────┐      ┌─────────────────────────────┐
      │    Exit    │      │  Continued web site browsing │
      │   [1824]   │      │            [2918]            │
      └────────────┘      └─────────────────────────────┘
                                      │
                          ┌───────────┴───────────┐
                          ▼                       ▼
                  ┌────────────┐     ┌──────────────────────────┐
                  │    Exit    │     │ Visited the "Contact Us" │
                  │   [2318]   │     │       page  [600]        │
                  └────────────┘     └──────────────────────────┘
                                                 │
                                    ┌────────────┴────────────┐
                                    ▼                         ▼
                            ┌────────────┐          ┌──────────────────┐
                            │    Exit    │          │ Sent an enquiry  │
                            │   [546]    │          │      [54]        │
                            └────────────┘          └──────────────────┘
```
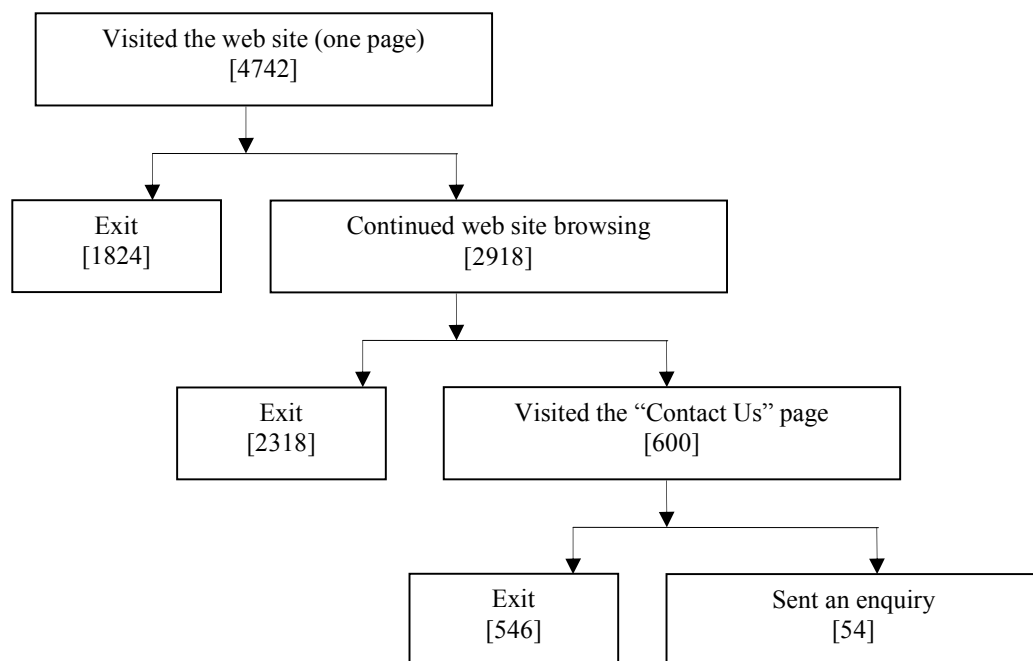
Diagram 1. The stages of visitors' decision.

A data set is collected from the log file for a web site of a small UK based web development agency and includes information about 4742 browsing sessions. We collected the next parameters for each browsing session:

− *hits* – a number of web pages attended by a visitor;
− *seconds* – time spent on the web site (in seconds). It's calculated as a difference between the first and the last page visit time, so time spent on the last page is not included.

- *dportfolio*, *dcost*, *dservices*, *dcasestudies*, and *dcontact_us* – dummy variables which is true if a specific page is visited during the browsing session (a page with previous jobs executed (portfolio), with price information, with information about services provided, with case studies and client's feedback, and with contact information respectively). The latter, *dcontact_us,* is used as a resulting choice variable for the first stage decision, but others – as explanatory variables.

- *denquiry* – a dummy variable which is true if a visitor sent an enquiry to the agency via a special form. The variable is used as a resulting for the second stage decision.

- *Monday-Sunday* – dummy variables for days of week.

- *Google*, *MSN*, and *Yahoo* – dummy variables for a session referrer (where did a visitor come from?). The web site is not well-positioned in other search engines, so visits from them are accidental and very rare.

- *kw_UK*, *kw_java*, *kw_london*, *kw_web* – dummy variables for keywords ("UK", "java", "London", and "web" respectively) used in search engines for finding the web site.

Results

Firstly we compared probit and logit discrete choice model specifications and didn't discover a significant difference between them (with the maximum likelihood ratio test) for both choice stages. We decided to use the logit model for further analysis.

The results of model parameters estimation ([3]) is presented in Table 1.

Table 1. Model coefficients estimation results.

Variable	Stage 1. A decision to visit a contact info page (dcontact_us)			Stage 2. A decision to send an enquiry (denquiry)		
	Coef.	*p-value*	*Marginal effects*[*]	*Coef.*	*p-value*	*Marginal effects*[*]
hits	0.239	0.000	0.03395	0.001	0.934	
seconds	-0.0001	0.358		0.000	0.009	0.00001
dportfolio	-0.415	0.003	-0.05114	-0.126	0.761	
dcost	0.059	0.658		1.089	0.005	0.04547
dservices	0.184	0.211		-0.315	0.446	

dcasestudies	-0.397	0.019	-0.04922	-0.253	0.556	
Google	0.510	0.000	0.08469	0.331	0.388	
MSN	1.207	0.027	0.23722	-15.792	0.997	
Yahoo	0.574	0.522		16.925	0.986	
Monday	-0.392	0.028	-0.04870	0.611	0.307	
Tuesday	0.164	0.307		1.250	0.015	0.05674
Thursday	-0.046	0.783		0.985	0.070	0.03899
Friday	-0.212	0.233		0.539	0.378	
Saturday	-0.675	0.002	-0.07600	0.053	0.951	
Sunday	-0.491	0.032	-0.05892	1.273	0.072	0.05846
kw_UK	0.409	0.030	0.06591	0.191	0.690	
kw_web	-0.188	0.173		0.175	0.653	
kw_london	0.785	0.002	0.14055	0.345	0.609	
kw_java	0.658	0.005	0.11391	-27.906	0.982	
Const	-2.515	0.000		-3.737	0.000	

[*] – marginal effects are calculated for sample average values of hits and seconds and zero values of dummy variables.

The table contains the estimated coefficients, p-values for them, and marginal effects calculated for significant variables.

Analysing the first stage model we can note the positive coefficient for *hits* variable which indicates that a visitor often carefully investigate the web site before going to a contact info page. The marginal effect for *hits* shows that each additional hit increase the probability of the contact page visiting by 0.03 or 3% (this value is calculated for average values of *hits* and *seconds* equal 4 and 207 respectively and zero values of dummy variables).

There are two significant page-specific variables – portfolio and case studies, both with negative signs. Both pages contain information about previous jobs executed by the agency, and the negative effects can be explained in two ways:

− shortcomings in the portfolio and case studies pages content (which is the true for the considered web site);

− a specific pattern of customer behaviour where he firstly investigates web sites of some agencies and only after that (in the next session) makes the decision about contacting.

Two of three search engines dummy variables are significant – *Google* and *MSN*. Both coefficients are positive, so referring the web site from these search engines is very good (for contact info page showing). *MSN* has a large marginal effect value (0.23), so the

probability of a visit of the contact info page for a person from MSN are 0.23 higher, than average.

Monday, *Saturday*, and *Sunday* variables have significant negative coefficients (*Wednesday* was selected as the dummy base and excluded from the model). If the *Saturday* and *Sunday* results are expected (usually clients don't really at work on weekends, just surfing for future contacts), but the *Monday* is something that requires additional explanations. We suppose that there are many internal jobs in companies-clients on Monday which make new contacts less probable.

Significant positive coefficients for "UK" and "London" (the place where the agency is situated) keywords dummies are expected – if a visitor decided the find an agency in his area, the contact info is highly important for him. Also we note a positive effect of "java" (a popular programming platform) keyword in the search line – this technology is the main agency's specialisation.

The second stage model has no so many significant coefficients. When a visitor already comes the contact information page, there are not many factors affecting his decision about sending an enquire. Time spent on the web site (*seconds*) has positive influence on the decision about sending an enquiry. A visit to the page with price information (*dcost*) is the only significant at this stage of the decision and it has an expected positive influence.

Conclusions

The main goal of this study was web site visitor pattern identification, discovering of factors, which influence on his decision about previewing the contact info and sending an enquiry to the agency.

The suggested model is based on the sequential modification of the popular discrete choice method. All parameters, included into the model, can be extracted from the log file, which is automatically collected on web servers. These facts make the suggested model open and easy-to-use and increase its practical meaning.

We estimated the model parameters for data for a small web development agency web site. Results were presented to the agency management and allow improving web site content, planning advertising pay-per-click campaigns with search engines and predicting server loading.

Bibliography

1. Ben-Akiva M, Lerman S. *Discrete Choice Analysis: Theory and Application to Travel Demand (Transportation Studies)*, Massachusetts: MIT Press, 1985, 384p.
2. An M. Y. *Econometric Analysis Of Sequential Discrete Choice Models*. Duke University, Working Paper No. 95-55, 1997, p.22.
3. Buis M.L. *SEQLOGIT: Stata module to fit a sequential logit model*, VU University Amsterdam, Software Manual, 2007, p.17.