ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2007, Vol. 36, No. 3

WEBSITE ACTIVITY ANALYSIS MODEL

Algirdas Noreika, Sigitas Drąsutis

Department of Multimedia Engineering, Kaunas University of Technology
Studentų St. 50, LT–51368 Kaunas, Lithuania

Abstract. Knowing website traffic flow and structure is essential in order to make website successful and better target visitors. In this article website traffic and structure analysis models in combination of intelligent methods are proposed and theoretical predictions are made on how and what factor changes in website structure will affect visitor click paths and overall website activity.

1. Introduction

Due to every day growth of internet services and e-commerce activities good website marketing understanding, strategies development and possibilities identification are essential to every commercial website.

Adequate solutions are necessary to help to identify possible website optimization techniques and models. To do so, we should know our customers – website visitors, to understand their needs. "As marketing guru Regis McKenna explains, 'it's about giving customers what they want, when, where and how they want it'" [1], or "The theory of customer relationship management can be summed up in one phrase: targeting the right offer to the right customer at the right time for the right price" [2]. This is only possible by analyzing website traffic flow, its origins as well as by knowing website content structure and how any changes to this structure could affect that flow.

In this article we propose website activity data analysis model. The main principle of this model is that we divide website analysis into two parts – website structure analysis model and website traffic analysis model. Our aim is to construct and formalize these models separately and then find a relation function between them based on intelligent methods. However, we ddescribe only models construction in this article leaving AI based function construct as a black box.

2. Website structure and traffic flow.

In a real world building new roads and setting up new landmarks controls traffic flow and rate. Changes in website organization will also change visitors flow dramatically. Although in website we do not need to optimize visitors rate so much, because here, contrary to the real world, we have no road permeability problems. On the other hand, for every website marketer, it is very useful to know how changes to website structure will affect its traffic flow, will it go to the one or another direction. But how website structure changes could affect website traffic flow is not as easy to predict as it may look at the first time.

At first let's focus on visitors activity concept and website traffic analysis. To create a website traffic analysis model, we must know where our visitors come from, how do they navigate through the website and how do they leave. All these visitors actions we call visitors activity, or website activity. Website activity measurement is highly dependent on data mining techniques so we must look at both ways: data model and data analysis model. Popular data mining techniques available today are discriminant analysis, decision tree induction, and neural networks including multiple linear regression [3]. We could use one or more of these techniques in our website activity analysis.

To know better what should we look for, to know our visitors, we have identified and described the main possible types of visitors activities (*see Table 1*).

Other important activities may be:

- Long time website browsing.
- Important page viewing.
- Specific link click to leave a site.

To finish each such activity, website visitor must reach it through a specific way, called click-path. For example:

Search > Product info > Product Reviews > Purchase.

These paths are one of the targets in our model. One of our goals, by analyzing these click-paths and mapping them to a website structure, is to predict how any changes to website structure – adding of new services or starting a new in-site marketing campaigns could affect website traffic flow. Another of possible click-paths usages could be to identify, analyze and

find near similar or similar paths – possible ways to change website structure and navigation to make it perform better. It is also known as website link structure optimization.

Table 1. Possible types of visitors activities

Activity	Description
Purchase	Purchases a product or service. Two main types of purchase: • single purchase • repeat purchase
Opt-in	Newsletter, e-magazine subscriptions.
Personal information submission	User registration. Additional information like income, interests, age, demographics is available.
RSS subscription	Company/Website news, products feeds subscription (interest in company/website activity, services and products)
Information printing	Prints out a valuable information.
Using "Send a friend function"	Sends a friend a valuable or interesting information about company or a website.
Download	Downloads document or application.
Search	Uses website search service to search for a specific information.

Table 2. Most popular websites traffic sources

Visitor type	Marketing method	Description
First time visitors	Search engines	Keywords based. Websites are indexed by search engine robots applications (called spiders) and added into search engines databases. Visitors reach these websites by searching for information they are interested in. SEO (Search Engine Optimization) is required to list websites effectively into SERPs (Search Engine Report Pages).
	PPC, PPI	Pay per Click, Pay per Impression systems. Keyword based systems. Highly targeted, website related keywords must be chosen. Relatively to the marketing method a constant amount of money must be paid for a website link (banner) click or display.
	Directories	Links to a website must be placed into a highly targeted category of a directory. Visitors come by browsing categories of directories they are interested in.
	Link exchange / Ad Exchange	Direct links and banners exchange between websites.
	Direct marketing	Direct marketing through traditional media: Television, magazines, newspapers, telephone marketing etc.
Return visitors	e-mail subscriptions	Visitors subscribe to website's newsletter or e-zine systems. They are always in touch with company's announcements, products and services.
	Fresh and attractive site content.	News, fresh and free information, products updates and downloads keep visitors back.

However click path analysis is not enough to analyze website activity. We must also evaluate website traffic sources to know where our visitors are coming from. These are as starting points for most of the click paths generated in a website. Table 2 shows a number of most popular website traffic sources [4].

As in the real world good road infrastructure is essential for logistics optimization, the same is for

internet website structure. Having a good website structure: website inner links, data blocks organization means that website visitors can easy navigate, search, purchase and perform other activity on a website. There are a number of models proposed regarding website structure analysis and optimization. Some models analyze inner link structure, while others use semantics - a keyword based connections between web pages.

Although website structure and traffic could be viewed as separate objects, they have a tight relation with each other. First, they share the same structure and website traffic flow directions depend on it (*see Figure 1*).

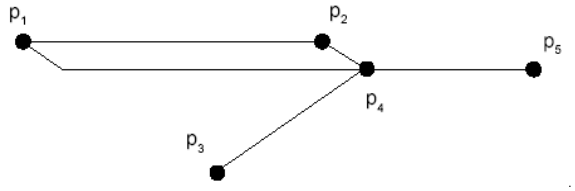

Figure 1. A graphical interpretation of basic website page structure

We can describe the whole website like a finite set of pages *P* assuming that the website has a number of web pages n, where each page is identified by its index *i*:

$$P = \{p_i\} = \{p_1, p_2, p_3, \ldots, p_n\}. \quad (1)$$

Second, as proposed in [5], there is a semantic connection between pages and also duration of visits affects visitors interest in one activity or another. We believe that by constructing the models of these two separate parts, website structure and traffic flow, and then by composing them into one we could create a website activity analysis and prediction mechanism the web marketers could use in developing their commercial websites.

In the next two sections we will describe our website structure and traffic analysis models in formal notation with some improvements compared to other models proposed in existing literature.

3. Website structure analysis model

The structure of website is very important. It defines the whole navigation system and content flow. Website could be imagined as a tree or a graph with nodes and links between them. Each node is a single page of a website. Each edge connecting different nodes is a link between pages. It is very important about how we create this structure, because website visitors traffic flow will be directed by it.

Most website structure models [6, 7] focus on website link structure and web pages semantic connections based on similarity of web pages content. The similarity in this case is measured using the cosine function between vectors, scaled according to the inverse-document-frequency paradigm, used in Information Retrieval. Most of these models analyze page content keywords only, leaving link and link title attribute keywords behind. We believe that link information in many cases is also very important so we improved our model with it. Firstly, to construct our website structure analysis model, we defined the four main factors included in this structure:

- *Pages.* Pages in our model are like nodes of a whole website structure. Pages are connected by links.
- *Page Content (content keyword vectors).* Each page has a meaning. This meaning can be defined as a set of separate keyword vectors constructed for each page.
- *Links.* Links in a website connect one page to another and in this way create a possibility for visitor navigation.
- *Links content (links keyword vectors).* Each link may also be described by keywords it contains. Link keyword vectors in conjunction with page keyword vectors create unique direction.

We have already defined the whole website as a finite set of pages in (1). Next we should define the relation between pages, page content, links and links content.

Let's start from the page content. Each page of a website can have a different types of content – text, images, media, etc. In our model we assume that we will analyze only textual part of the page due to complexity of analysis of other types. Text analysis is not so easy as it may look at the first moment but, fortunately, it is widely investigated in many computer science related papers, so we will solve this problem in one way or another. Mainly, we should define the set of keyword vectors, a short description, of a page we are analyzing. So every page will have a finite set of keyword vectors k related to it:

$$\begin{aligned} K_1 &= \{k_1, k_2, k_3, \ldots, k_{m1}\} \\ K_2 &= \{k_1, k_2, k_3, \ldots, k_{m2}\} \\ &\ldots \\ K_n &= \{k_1, k_2, k_3, \ldots, k_{mn}\}. \end{aligned} \quad (2)$$

$$PK_i = \{K_j\}. \quad (3)$$

where $j = 1, n$.

From here keyword vector set in a page can be defined:

$$P = \{p_i\{PK_i\}\}. \quad (4)$$

where *K* is different set of keyword vectors for each page.

Links definition is very similar to pages definition:

$$\begin{aligned} L_1 &= \{l_1, l_2, l_3 \ldots l_{m_{links}1}\} \\ L_2 &= \{l_1, l_2, l_3 \ldots l_{m_{links}2}\} \\ &\ldots \\ L_n &= \{l_1, l_2, l_3 \ldots l_{m_{links}n_{links}}\}. \end{aligned} \quad (5)$$

$$PL_i = \{L_{j_{links}}\}. \quad (6)$$

where $j_{links} = 1, n_{links}$.

A finite set of links keywords vectors is as follows:

$$C_1 = \{c_1, c_2, c_3 \ldots c_{m_{kwd}1}\}$$
$$C_2 = \{c_1, c_2, c_3 \ldots c_{m_{kwd}2}\} \quad (7)$$
$$\ldots$$
$$C_n = \{c_1, c_2, c_3 \ldots c_{m_{kwd}n_{kwd}}\}.$$

$$LC = \{C_{j_{lkwd}}\}. \quad (8)$$

where $j_{kwd} = 1, n_{lkwd}$.

From here a keyword vector set in a link is defined:

$$L_{j_{links}} = \{l_{i_{links}j_{links}}\{LC\}\}. \quad (9)$$

A combined (2) and (4) for each page:

$$P = \{p_i\{PK_i, PL_i\}\} \quad (10)$$

It is important to remember that both *PK* and *PL* are different sets for different pages. The last equation defines our website structure model as seen in Figure 2.

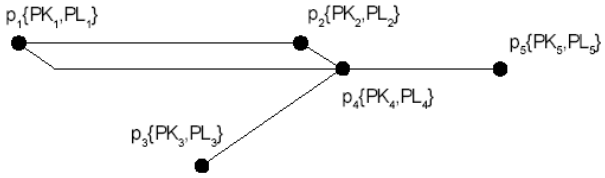

Figure 2. A graphical interpretation of website structure model

4. Traffic analysis model

Website traffic analysis is one of the ways to know what our visitors are looking for. One of the most popular methods of website traffic analysis is click-stream data analysis. "Analyzing click-stream data is becoming the most important activity for e-business. Click-stream analysis can reveal usage patterns on the company's web site and give a highly improved understanding of customer behavior. This understanding can then be utilized for improved customer satisfaction with the website and the company in general, yielding a huge business advantage" [8]. There are already a number of methods and models proposed regarding to website traffic analysis [8-10] like time-stamp based click-stream analysis or data packets analysis model for the intranet [11]. In this section we should use a number of data mining techniques to acquire the right information in right time. After studying a number of website traffic analysis models, we have defined three main factors for traffic analysis and suggested a fourth factor which defines the origin of traffic and describes where traffic comes from. This factor is very important in website activity analysis. We have described all these four factors influencing website traffic flow characteristics below:

- *Click-stream.* Click-stream or click-path defines a whole click-path generated by a single visitor while visiting a website from the entry point to the exit point.
- *Timing.* Timing defines the duration of the whole visit and the duration of each page visit also. In most cases this reflects visitors interest in a website or any single page of it.
- *Weight.* Weight or traffic rate shows how website traffic is distributed across its structure. This factor also reflects visitors interest in some parts of a website.
- *Source*. Traffic source shows where visitor came from.

Let's assume that website had a number of visits in some period of time. A whole set of visitors could be described like:

$$X = \{x_{i_{visitors}}\} = \{x_1, x_2, x_3, \cdots, x_{n_{visitors}}\}. \quad (11)$$

Looking back at the definition of website structure (1) and Figure 1 a single click-path could be like:

$$p_1 \rightarrow p_2 \rightarrow p_4 \rightarrow p_5 . \quad (12)$$

A duration of visit for a single page could be described like:

$$\Delta t_j = T_{j_{exit}} - T_{j_{entry}} , \quad (13)$$

where j is a page index in a click-stream, $T_{j_{entry}}$ is page entry time-stamp and $T_{j_{exit}}$ is page exit time-stamp. The whole duration of a visit for a single visitor can be easy calculated:

$$t_{visit} = \sum_{i_{clickstream}=1 \ldots n_{click}} \Delta t_{i_{clickstream}} . \quad (14)$$

The traffic weight w can be defined as visitors rate n_i through a single page during the website traffic analysis period:

$$N = \{n_{i_{rate}}\} = \{n_1, n_2, n_3, \ldots, n_{n_{rate}}\}. \quad (15)$$

$$w_i = \frac{n_i}{\max(N)} . \quad (16)$$

Traffic flow s also depends on traffic source type and the amount of traffic s_i it sends:

$$S = \{s_{i_{rsourse}}\} = \{s_1, s_2, s_3, \ldots, s_{n_{sourse}}\}. \quad (17)$$

$$s_i = \frac{s_i}{\max(S)} . \quad (18)$$

The whole set of factors defined above affects traffic flow and its direction on each page $p_i\{\Delta t_i, w_i, s_i\}$. The final graphical interpretation of traffic analysis model is shown in Figure 3.

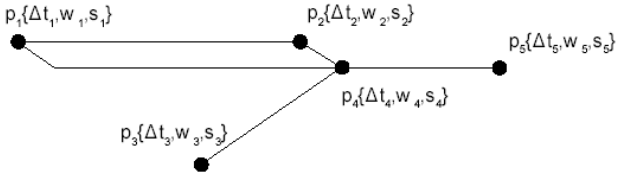

Figure 3. A graphical interpretation of traffic analysis model

5. Website activity analysis model

In the previous sections we have defined website structure and traffic analysis models. We believe that after finding the relation functions between them we could model website traffic flow by changing website structure or to find possible website structure improvements by modeling its traffic. As a connecting part for these models (Figure 4), we plan to use AI based methods like neural networks or fuzzy logic constraints, which are well known as nonlinear functions for problem solving, to help us to find a relation how changes in one model could affect the other. Surveys covering the whole range of business applications of neural networks can be found in [12-15]. However, definition of AI functions is outside the scope of this article and we left it for future investigation.

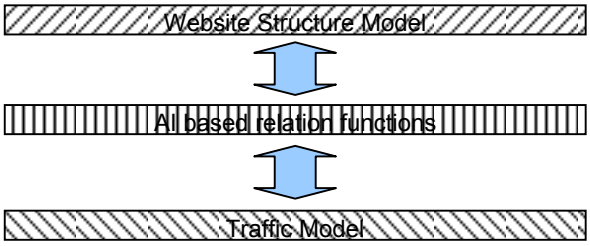

Figure 4. Traffic and Structure analysis models connected with AI based nonlinear functions (black box)

6. Conclusions

Currently we have defined website activity model only. Our future work will be to define a connection mechanism between two parts of this model - website structure analysis model and website traffic analysis model and test it with various sets of data. We have studied a number of articles which propose different methods for website structure, traffic and activity analysis. Most of them focus on website link structure optimization, web traffic analysis or visitors segmentation. However we have found no method or model for website traffic prediction in relation to structure change of a website. We have studied website traffic and structure models from these articles separately and suggested some improvements. We believe that by creating a composite model we could create a method for website activity analysis and prediction.

Our future aim is not only to prove that connection between these two models exist, but also how changes in one model affect the other. It is easy to see that here we have to deal with a number of nonlinear relations. We may change link keyword, but we do not know how website traffic will be affected by this change, because traffic is also dependable on a number of other parameters. One of the ways to try to solve these nonlinear functions is to use already widely accepted nonlinear methods like neural networks. It is well known that neural networks are widely used in business applications for prediction, classification and other problems.

References

[1] **P. Fingar, H. Kumar, T. Sharma.** Enterprise E-Commerce. *Megan-Kiffer Press*, 2000, ISBN: 0929652118.

[2] **M.J.A. Berry, G. Linoff.** Mastering Data Mining: The Art and Science of Customer Relationship Management. *Wiley Computer Publishing*, 2000, ISBN 0-471-33123-6.

[3] **W.E. Spangler, J.H. May, L.G. Vargas.** Choosing data-mining methods for multiple classification: Representation and performance measurement implications for decision support. *Journal of Management Information Systems*, 1999, 16(1), 37-62.

[4] **A. Noreika.** Application of Intelligent Methods in Commercial Website Marketing Strategies Development. *Information Technology And Control, Kaunas, Technologija*, 2005, *Vol*.34, *No*.2, 140 - 144.

[5] **M. Kitajima, N. Kariya, H. Takagi, Y. Zhang.** Evaluation of Website Usability Using Markov Chains and Latent Semantic Analysis. *IEICE TRANS. COMMUN., Vol*. E88–B, *No*.4, *April* 2005.

[6] **B. Poblete, R.B. Yates.** A Content and Structure Website Mining Model. *ACM* 1595933329/06/0005, 2006.

[7] **B. Zhou, J. Chen, J. Shi, H. Zhang, Q. Wu.** Website Link Structure Evaluation and Improvement Based on User Visiting Patterns. *ACM* ISBN 1-59113-420-7/01/0008, 2001.

[8] **J. Andersen, R.S. Larsen, A. Giversen, T.B. Pedersen, A. H. Jensen, J. Skyt.** Analysing Clickstreams Using Subsessions. *ACM*, 2000.

[9] **A. Agrawal, J. Basak, V. Jain, R. Kothari, M. Kumar, P. A. Mittal, N. Modani, K. Ravikumar, Y. Sabharwal, R. Sureka.** Online marketing research. *IBM Journal of Research and Development, Vol*.48. *No*. 5/6, 2004.

[10] **X. Wanga, A. Abrahamb, K.A. Smith.** Intelligent web traffic mining and analysis. *Journal of Network and Computer Applications* 28, 2005.

[11] **P. Defibaugh-Chavez, S. Mukkamala, A.H. Sung.** Website Visitor Classification Using Machine Learning. *IEEE Fourth International Conference on Hybrid Intelligent Systems* 2/04, 2004.

[12] **A. Refenes, A.N. Burges, Y. Bentz.** Neural Networks in Financial Engineering: A study in methodology. *IEEE Transactions on Neural Networks* 8(6), 1997.

[13] **A. Vellido, P.J.G.Lisboa, J.Vaughan.** Neural networks in business: a survey of applications (1992 – 1998). *Expert Systems with Applications* 17(1): 5170, 1999.

[14] **B.K. Wong, T.A Bodnovich, Y. Selvi.** Neural network applications in business: A review and analysis of the literature (1988-95). *Decision Support Systems* 19: 301320, 1997.

[15] **G. Zhang, B.E. Patuwo, Y.M. Hu.** Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting* 14(1): 3562, 1998.

Received June 2007.

DOI: 10.5755/j01.itc.36.3.11884