Mathematical Problems of Computer Science 41, 93--102, 2014.

A Probabilistic Model of User Navigation in a Digital Interactive Publication

Vahram H. Darbinyan

State Engineering University of Armenia (Polytechnic)
e-mail: dvahram@gmail.com

Abstract

A probabilistic model of user navigation in a digital interactive publication is considered. Digital interactive publication is a paged media where user can navigate either by flipping pages back and forth or using a link for navigation to another page. User's navigation by links is analyzed in detail. For that case a special weight is assigned to any navigation from the current page to another page. A method of calculating these weights is proposed basing on a publication type and structure. Considerations include both server-side and client-side analysis. A mapping of the introduced weights to probabilities of navigation to each page is proposed also. These probabilities allow predicting user behavior during the paged media navigation.

Experiments with a representative sample set of digital publications have been performed to illustrate the validity of the suggested model.

Keywords: Digital interactive publication, User behavior, User navigation, Navigation model, Navigation prediction, Server-side analysis, Client-side analysis.

1. Introduction

User navigation is widely used in modern web based interactive technologies. Many websites and web based services have issues that users become disoriented: they lose sense of location and direction, so service providers try to come up with best practices of website structure. User's behavior and action prediction is also practiced in attempts of enhancing user experience [1, 2].

Prediction of the nextaction can be very beneficial for many web-based services and it can allow preparing and caching the expected next requested page, starting transfer of a video ortaking other actions that can improve the user performance overall.

When visiting a website the user's behavior is unpredictable in general. The visitor can follow any link from the menu. There are special tools that allow analyzing the user's behavior in website [3]. They provide statistical and analytical data that allow building websites which are more suited for a particular goal.But still generalizing the visitor's behavior in websites is a complicated problem and it's hard to create a general applicable model to benefit from.

Digital interactive publications (DIP) are unique media [4] where users' behavior is more predictable in comparison with a regular website. While people browsing the Internet usuallyare in pursuit of some specific information,the objectivewhen reading a digital magazine usually is

leisure and entertainment. In Internet pages a visitor usually navigates from page to another by links seeking the information he needs or is interested in. Unless there is a certain "wizard" interface with "next" and "previous" buttons, the user behavior is random and unique for each website [1, 5, 6].

Taking into account the mentioned predictability of a user behavior when reading a digital publication a probabilistic model of user's navigation can be created. Such a model would allow estimating probability of user's next action via the built model.

Using statistical data for the user's action prediction is generally easier and more accurate. At the same time a diversity of behaviors for different users when reading the same publication is not big and, therefore, the decisions made on statistical data are usually accurate if some other aspects such as used device are considered too.

The main drawback of statistics-based decision is the availability of data. A publication that has been published for a while and has been read multiple times leads to corresponding data which are usually available, while new publications lack such statistics. In that case we can't benefit from being able to predict users' next action for enhancing the user performance as well as for making other decisions.

The above described unavailability of data is the case when such a model can be extremely helpful as it would allow predicting the user's behavior without using statistical data.

There exist user behavior models based on hyperlink analysis that allow predicting user's navigation via navigation history and link structures [1, 7, 8]. In [8] a navigation behavior pattern is proposed, which can be used to predict visitor's navigation through the web. However, the existing models are designed for websites and do not incorporate important aspects of the investigated object such as a publication structure or type, which are the key for predicting user behavior in the case of reading a DIP.

A navigation model defines the content, structure, and metadata of the navigation. It specifies the items to include and the hierarchy of those items. Navigation models typically include the following resources [9]:

- Pages (individual pages and page hierarchies)
- External links
- Content (individual content items or the results of a content query)
- Other navigation models

Modeling of user navigation via page flipping and links will be concentrated on further.

Section 2 considers how the publication type affects user navigation. An influence of user devices on the navigation will be discussed too. Based on these factors weights will be calculated for each navigational event: flipping a page and navigating via a link.

In section 3 the way of calculation of the proposed weights is described. Weights are calculated for each page depending on page's displacement relative to the current page (e.g., if the page is next or previous page for the current one) and surface of links from current to the considered page. Mapping of weights to probabilities and building of a probabilistic model is described too.

Section 4 incorporates client-side data into the built probabilistic model. Using the client-side information will enhance the accuracy of calculations in cases when navigation prediction is performed on a user device.

Section 5 describes experiments performed with a sample set of digital publications. Data illustrating the validity of the suggested model are adduced there.

The conclusion summarizes the obtained results and outlines further investigations.

2. Server-side Factors that Affect User Navigation

Many factors affect users' behavior, e.g.,a structure of the publication, its content, interactive elements like videos and links and many others. These factors vary in their effect and those that have most significant effect are considered below. These data are available at the server-side and are intended for prediction of users' navigation actions via server.

2.1. Types of Publications

Types of digital publications differ in waysusers read them. Behavior of a userreading a fiction book differs from behavior when a scientific/research article or a magazine is read [10]. Thereforethe model to be built should take into account this difference.

To make the consideration observable we suggest presenting the existing variety of publication types: magazines, catalogues, photo albums, articles, journals, manuals, newspapers, brochures, etc. distributed into the following threebig groups.

Sequential content: The group will contain publication types that have mostly monolithic and sequential content, like fiction books where the content is a single story and is supposed to be read sequentially. Here the navigation is sequential and links are rarely used. Bookmarks are used to continue reading, but from the moment reading is restarted from a given page the navigation once again becomes sequential.

Discrete content: This group is comprised of catalogues, manuals and publications of other similar types. The navigation here is mostly non sequential, the content consists of separate articles, sections and titles that are rarely read altogether. Few users will read a manual completely at once, while many users will be referencing it periodically exploiting the links in table of content for navigation.

Mixed content: Publications like newspapers, magazines, and journals are forming the group, where depending on a use case, the content can be read either with a sequential navigation, e.g. flipping through the content or using the links.

2.2. Device Type

Depending on a user device and the size of the screen the behavior of the user might be totally different [11, 12]. Here too user navigation in a DIP can be either sequential or link-oriented depending on the device and screen size. In case of sequential behavior there is a greater chance the user will be scrolling through pages and there are not less cases when the navigation is link-oriented.

The following way of describing this difference in the model is suggested. Chance of a link to be clicked is described by a pair of values: L and C. L stands for the coefficient of clicking a link on any page other than the cover page. The chance of clicking a link on the cover page is bigger, so C will be used to denote that case. Heuristically calculated values are used for these coefficients.

The values are calculated basing on data gathered from more than 250000 digital publications and more than 100000 publishers. The gathered data is analyzed from the navigation point of view in terms of page flipping and link usage. Initial statistical processing

was performed using live Digital Interactive Publication Service. Over 80 million records of user navigation were analyzed; more than 30 million among them were in publications considered as having sequential navigation and more than 50 million - navigations in publications considered as having link-oriented navigation. In more detail this distribution is adduced below.

Sequential navigation: Here over 4.7 million navigations were from cover page and almost 25.3 million navigations were from non-cover pages. Among over 4.7 million navigations from cover page 1.7 million were using links, which brings to ~36% of the total number of navigations. Among 25.3 million navigations from pages other than cover page, over 5.2 million were using links, which is ~21% of the total number of navigations.

Link-oriented navigation: Here over 43,3 million were navigations from pages that are not cover page and 6.8 million were navigations from cover pages. Among them, ~46% (19.9 million) of navigations from non-cover pages use links. ~52% (3.5 million) of navigations from cover page use links.

Thus in the case of sequential navigation the values of coefficients can be taken as L=0.21 and C=0.36 [12].

When user navigation is link-oriented the chance of user clicking a link is higher compared to sequential navigation. The link-oriented navigation is common for devices with small displays or slow Internet connection where users choose to use links instead of flipping from page to page [11]. In this case coefficients of link navigation are higher: L=0.46 and C=0.52.

To express this formally a notation L(x,Nm) is introduced for link navigation factor, where *x* is a number of the current page and Nm is the navigation mode: 1 – corresponds to sequential, 2 – to link-oriented.

$$Values\ of\ L(x, Nm) for\ the\ considered\ cases:$$

	$Nm = 1$	$Nm = 2$
$x = 1$	0.36	0.52
$x \neq 1$	0.21	0.46

2.2. Structure of Publication

Other important aspect that affects users' behavior when reading a digital publication is a structure of the publication itself, more specifically the presence of links from one page to another. If there is a link on page *x* to page *y* the chance of page *y* being visited when the user is on page *x* is bigger. If the page *x* is a cover page, the chances are even bigger.

Another factor to be taken into account while modeling the navigation from page *x* to page *y* is not only the mere presence of a link from page *x* to page *y*, but the size and location of the link on the page. Links that are larger are more likely to be used compared to smaller ones. Links can also be placed over the whole page which makes the probability of them being used even higher (sometimes by mistake).

Assuming there is a number of links on page *x*, let's name that set of link objects L^x. It is either empty orcan be expressed as:

$$L^x = \{l_1^x, \dots\ l_{n_x}^x\},$$

where n_x is number of links on the page *x*. Let *y* be a page lead by some links from the set L^x. A subset of L^xcontaining all links that lead to *y* is denoted as L_y^x:

$$L_y^x = \{l_{y_1}^x, \dots\ l_{y_m}^x\},$$

where *m* is number of links on the page *x* leading to the page *y*.

The chance of navigating from page *x* to page *y* using links is correlated to surface of links on the page *x* that lead to the page *y*. *Fs(x,y)*notation is introduced to denote that chance.

For calculation of *Fs(x,y)* the sum of surfaces *S(x,y)* for all links to the page *y* from the page *x* has to be calculated.

$$S(x,y) = \sum_{i=1}^{m} sur(l_{y_i}^x),$$

where *i* is the number of the linkobject $l_{y_i}^x$from the set L_y^x and $sur(l_{y_i}^x)$ is the surface of that linkobject. Also a total sum of link object surfaces is required for *Fs(x,y)* calculation. *S(x)* stands for that sum.

$$S(x) = \sum_{j=1}^{n_x} sur\left(l_j^x\right).$$

where $sur\left(l_j^x\right)$ is surface of link object $l_j^x$from set L^x. Finally, *Fs(x,y)*is calculated using the following formula:

$$Fs(x,y) = \begin{cases} \dfrac{S(x,y)}{S(x)}, & S(x) > 0, \\ 0, & S(x) = 0. \end{cases}$$

3. Creating the Model of User Navigation

It implies from the considerations above that in our model the chance of user navigation from page *x* to page *y* depends on the following factors:

- Publication type,
- Number of the current page,
- Does the current page coincide with the previous page?
- Set of links from the current page to the target page (number of links and their surface)

The way of calculating this chance is adduced below.

3.1. Calculating Weights for Pages

When the user is on page *x*, three following actions can be taken in terms of navigation:

1. Navigate to next page
2. Navigate to previous page
3. Use a link to navigate to some other page

For predicting which page is going to be navigated by the user while being on the page *x*, it is suggested to use special weights for all pages to be navigated. These weights should be calculated beforehand.

The weight of a given page is correlated to the chance of being navigated. Weights are calculated basing on page's dislocation relative to the current page and also links from the current page to the considered one. Three measures are introduced below: w_n, w_p and w_l.

w_n is introduced to indicate if the number *y* of the next page immediately succeeds the number *x* of the current page. w_n will have value *1* if the condition of being the immediately succeeding numberis true and *0* in cases when it is not true.

To indicate the fact that the number y of a givenpage is the immediately preceding the number *x*of the current page, w_p is introduced. Similarly, $w_p = 1$ when $x = y + 1$ and $w_p = 0$ in all other cases.For the value 1, w_n and w_p are mutually exclusiveas any page can't be simultaneously previous and next page for a given page.Thus w_n and w_p can't have value 1 at the same time.Presence of links on the page *x* to the page *y*should increasethe weight of the page *y*. The measure w_l is to reflect that:

$$w_l = Fs(x, y) * L(x, Nm).$$

The total weight for the page *y*is calculated by summing the values of considered measures:

$$W_y = k_n w_n + k_p w_p + k_l w_l,$$

where k_n, k_p and k_l are coefficients of influence for the corresponding measures. They are calculated heuristically.

Initial calculation of k_n, k_p and k_l. Analysis of gathered statistical data on navigation actions is performed. From the same ~80 million navigation records mentioned above, almost 50 million records were navigations via page flipping. Among those over 42 million records were navigations to the next page, which is ~53% of total 80 million navigations. Navigations to previous page are counted in more than 7 million records: ~9% of the total. Thus for k_n and k_p values like 53 and 9 can be used. As $w_l$already includes the chance$L(x, Nm)$ of a link to be used, a value 100 is used for k_l.

Tuning of k_n, k_p and k_l in the process of use. Initially calculated values should be adjusted during their use. If according to the built model abehavior of a user is described asusing a link for navigation, but instead a page was just flipped to the next page, the coefficient of page flipping (k_n) should be increased and the coefficient of using a linkk_l should be reduced. In the casethe prediction was correct, no action should be taken against the coefficients. Per our experiments after a sufficient number of publication readings the coefficients are reaching some stable values.

So it is assumed further that each page has a calculated weight for navigation.

3.2. Mapping Weights to Probabilities

Sum of weights for each page of a given publication is needed to calculate probabilities for pages. $W_T$will be used to denote that sum of page weights. It is calculated by the formula below:

$$W_T = \sum_{i=1}^{p, i \neq x} W_i,$$

where p is the number of pages of the publication that is being analyzed. $W_i$stands for weight of a page with the number i.

$p_y(x)$ notation is introduced to express the probability of page y being navigated while the user is on page x. $p_y(x)$ is calculated using the following formula:

$$p_y(x) = {W_y}/{W_T},$$

where W_y is the weight of the page y andW_T is sum of weights of all pages.

Based on the calculated probabilities, pages are sorted by their probability values in descending order. The page with highest probability is predicted to be the next in navigation sequence.

4. Using Client-side Information

The model proposed can be complemented with the client-side data to enhance the accuracy. Though using the client-side information limits the usage of the model to only client side, at the same time it provides higher accuracy.

Some devices cannot fit the whole page of the publication on the screen, or they require zooming in to provide acceptable reading performance [11]. In the case of zoomed page or scrolling, not all links on current page are visible to the user, thus the chance of usage for some of them is reduced. Considering that on many devices where page zooming or scrolling is required, the user's navigation is link-oriented the accurate calculation of probability of navigating via a link is paramount.

In case of zoomed pages or scrolling, when only a part of the page is visible at a moment, instead of calculating the surface of all links to a certain page from the current one, surface of visible links should be calculated. So a L^{x_v} set of visible links should be considered instead of $L^x$and $L_y^{x_v}$ set of visible links leading to page y should substitute L_y^x:

$$L^{x_v} = \{l_1^{x_v}, \dots l_n^{x_v}\}, \quad L_y^{x_v} = \{l_{y_1}^{x_v}, \dots l_{y_m}^{x_v}\}, \quad L_y^{x_v} \subseteq L^{x_v}.$$

Accordingly $S(x, y)$ will be calculated by the following formula:

$$S(x, y) = \sum_{i=1}^{m} sur(l_{y_i}^{x_v}),$$

where $sur(l_{y_i}^{x_v})$ is surface of visible link object$l_{y_i}^{x_v}$from set $L_y^{x_v}$.This case also implies comparing $S(x, y)$to visible part of the screen instead of the surface of the whole page. $S_v(x)$denotes the visible surface of the page and is calculated similarly to the above. Thus $Fs(x, y)$ is calculated according to thebelow:

$$Fs(x,y) = \begin{cases} \frac{S(x,y)}{S_v(x)}, & S_v(x) > 0, \\ 0, & S_v(x) = 0. \end{cases}$$

5. Evaluating the Validity of the Suggested Model

Experiments have been conducted on a sample set of digital interactive publications. The set consists of publications of all groups: publications with sequential content, discrete and mixed contents. The suggested model is implemented in digital interactive publication client-side reader software. The client software is provided with the necessary publication metadata from the server side: number of publication pages, type of the publication, weight factors. The software is also "aware" of the user device it is being run on. Base on the available data and user's current page user's navigation is predicted. According the prediction resources of the predicted page are being pre-loaded, to reduce wait time for the resources and enhance user experience while reading the digital interactive publication. After the user navigates to any other page, both the actual destination page number and predicted page number are sent to server. These values are stored in server database. Based on the stored values accuracy of the suggested model is evaluated.

The experimental set of publications consisted of 27000 publications, published between Dec 1, 2013 and Mar 1, 2014. This set includes:

- 23324 Magazines
- 531 Catalogues
- 120 Photo Albums
- 380 E-Books
- 17 E-Cards
- 146 Articles
- 27 Essays
- 189 Journals
- 364 Manuals
- 553 Newspapers
- 239 Portfolios
- 897 Reports
- 665 Brochures
- 37 Comics

During the period of model evaluation 24 million predictions were made. The accuracy of predictions made using the suggested model ranged between 67 to 79% depending on publication type.

6. Conclusion

A probabilistic model for user navigation was proposed that allows predicting users' next action while reading a digital interactive publication. The model is used for action prediction in digital interactive publication services to predict user's navigation and to pre-load the page user

is most probably about to visit, thus enhancing the user experience. The model is also used to substitute statistical data when such is not available.

The further development of the suggested model is connectedwith adopting the model to behavior of each user instead of having a common model for all readers of the publication. Also elements of machine learning methods will substitute gradually the heuristics based calculations.

References

[1] E. Herder, "Modeling user navigation", *Lecture Notes in Computer Science*, vol. 2702, pp. 417--419, 2003.

[2] N. Wunderlich, F. Wangenheim and M. Jo Bitner, "High tech and high touch: a framework for understanding user attitudes and behaviors related to smart interactive services", *Journal of Service Research*, February 2014.

[3] M. Eirinaki, M. Vazirgiannis, "Web mining for web personalization", *ACM Transactions on Internet Technology,* vol. 3, no. 1, pp. 1--27, Feb 2003.

[4] W. Rosenblatt, S. Mooney and W. Trippe, *Digital Interactive Publications (DIP) are Unique Media*, John Wiley & Sons, Inc., New York, NY, USA, 2001.

[5] V. H. Darbinyan and R. H. Vardanyan, "Detecting DDoS attacks on digital interactive publication hosting services", *Proceedings of SEUA Series "Information technologies, Electronics, Radio Engineering"*, Yerevan, Armenia, vol. 15, pp. 20-25, 2012.

[6] R. H. Vardanyan, "An adaptive method for visualization of the interactive multimedia publications", *Proceedings of the Conference Computer Science and Information Technologies*, Yerevan, Armenia, pp. 332--334, 2011.

[7] W. T. Fu and P. Pirolli, "SNIF-ACT: A cognitive model of user navigation on the world wide web", *Human-Computer Interaction*, vol. 22, no. 4, pp. 355--412, 2007.

[8] J. Borges and M. Levene, "Data mining of user navigation patterns", *Lecture Notes in Computer Science*, vol. 1836, pp. 92-112, 2000.

[9] D. Cohn and T. Hofmann, "The missing link – a probabilistic model of document content and hypermedia connectivity", *Advances in Neural Information Processing Systems13 - Proceedings of the 2000 Conference*, pp. 430--436, 2000.

[10] R. Chen, A. Rose, B.B. Bederson, "How people read books online: mining and visualizing web logs for use Information", *Lecture Notes in Computer Science*, vol. 5714, pp. 364--369, 2009.

[11] R.H. Varda*nyan,* "Adapting the interactive multimedia publication content to mobile devices", *Proceedings of the Yerevan State University. Series*, Yerevan, Armenia, vol. 1, pp. 51--54, 2013.

[12] V. H. Darbinyan, "Analysis of user behavior when reading digital interactive magazine online on various devices", *Proceedings of the Conference Computer Science and Information Technologies*, Sep. 23-27, Yerevan, Armenia, pp. 445--446, 2013.

Submitted 25.12.2013, accepted 26.02.2014.

Թվային ինտերակտիվ հրապարակման մեջ օգտագործողի ուղորդման հավանականային մոդել

Վ. Դարբինյան

Ամփոփում

Հոդվածում նկարագրված է թվային ինտերակտիվ հրապարակումներում օգտագործողի ուղորդման հավանականային մոդել: Թվային ինտերակտիվ հրապարակումը էջային կառուցվածքով միջավայր է, որտեղ օգտագործողի ուղորդումը հնարավոր է կամ էջից էջ առաջ կամ ետ թերթելով, կամ հղումներ օգտագործելով: Հոդվածում վերլուծված է օգտագործողի վարքագիծը: Ամեն էջ ուղորդման տեղի ունենալուն վերագրված է որոշակի կշիռ, որը հաշվարկվում է հրապարակման տիպից և կառուցվածքից ելնելով: Կշիռներից կախված, ամեն էջի համար հաշվարկված է այդ էջ ուղորդման հավանականությունը:

Вероятностная модель навигации пользователя в цифровых интерактивных публикациях

В. Дарбинян

Аннотация

В данной статье предложена вероятностная модель навигации пользователя в цифровых интерактивных публикациях. Цифровые интерактивные публикации являются средой со страничной организацией, где навигация пользователя происходит или по средствам листания страниц, или же использованием ссылок. Приведен анализ поведения пользователя. Навигации с текущей страницы публикации на каждую другую страницу присваевается вес. Вес расчитывается исходя из типа публикации и ее структуры. Вероятности навигации на каждую страницу расчитываются исходя из весов.