Tomasz Zdziebko

A procedure for learning customer preferences in e-commerce websites

Folia Oeconomica Stetinensia 11(19)/1, 33-42

2012

VERSITA

Folia Oeconomica Stetinensia
DOI: 10.2478/v10031-012-0024-7

Wydział Nauk Ekonomicznych i Zarządzania
Uniwersytetu Szczecińskiego

A PROCEDURE FOR LEARNING CUSTOMER PREFERENCES IN E-COMMERCE WEBSITES

Tomasz Zdziebko, Ph.D.

University of Szczecin
Faculty of Economics and Management
Institute of IT in Management
Mickiewicza 64, 71-101 Szczecin, Poland
E-mail: tomasz.zdziebko@wneiz.pl

Received 26 July 2012, Accepted 26 October 2012

Abstract

High quality knowledge of users' preferences are of high value for every e-commerce website. A user's interest in a web page can be estimated by observing his or her behaviors. Implicit methods are less accurate than explicit methods, but an implicit observation is done without interruption of having to give ratings for viewed items. This article proposes a new procedure for obtaining e-commerce customer's behavior data and learning about their preferences from them. The main contribution of this procedure is an extension for FireFox browser which allows running a study in a user's natural environment. The extension created for the purpose of this study allows of monitoring a variety of events which are used for calculating implicit interest indicators. Another contribution of this study is the proposal of 5 new implicit indicators, four of which are designed especially for e-commerce websites.

Keywords: implicit feedback, e-commerce, human-computer interaction, preference modeling.

JEL classification: L81, C63, C81.

Introduction

From the beginning business owners strive for methods which will allow them to learn customer's needs and preferences. High quality knowledge about users' preferences is of great value for every company. It can be utilized in order to satisfy customers' needs more accurately. In the web environment this knowledge can be used to personalize an offer by improving the quality of recommendations.

Determining a user's interest can be performed explicitly by asking the user directly, or implicitly by observing the user's behavior. Implicit measures are generally less accurate than the explicit ones[1], but they are available in large quantities and can be acquired without any extra time or effort from the user. Requiring the users to explicitly rate items disrupts their normal reading and browsing behavior[2]. Moreover, unobtrusive monitoring of the users allows them to focus on a task at hand without any interruption caused by having to give ratings for the viewed items. Another advantage of implicit feedback over the explicit one is that it is difficult to motivate them to continuously give explicit ratings[3] even when the benefits are obvious (i.e., personalized recommendations).

Implicit feedback in a website area can be acquired from different sources on different sides of the interaction process, such as[4]: the server side, the proxy side, the client side. Historically, the first data source were server logs the main purpose of which was a diagnostic tool. Due to their limitations server logs store only information about every page request to particular websites installed on a server. The role of Proxy logs is the same as server logs, but they can store information about every access coming from users of this server. In contrast to the server and proxy side, the logging user's behavior at the client side, allows of a more complete grasp of the user behavior. Highly accurate logging at the client side is possible thanks to very good support for the Document Object Model across modern browsers. Observing users at a client side allows for high granularity monitoring of user's actions e.g. moving a mouse, scrolling contents, clicking, selecting contents, copying contents, saving pages, printing pages, key input. Wide range of monitored actions allows online shops to gain extensive amount of data about the user behavior. By using these data we can reason about the user's preferences more precisely and reliably.

Most researchers in the field of implicit feedback indicators focused mainly on evaluating the user behavior at an unlimited scope of websites[5]. Participants of these studies usually visited websites from many different categories (e.g. information portal, blogs, company websites). One of the drawbacks of this approach is the fact that the user behavior may depend on the type

of a site as well as it can be different on particular websites. These may be caused by differing users' goals on websites from various categories. Another factor influencing the user behavior may be caused be layout and functionality which is different for every website. When studying the literature the author encountered very few studies concerning the evaluation of implicit feedback in the e-commerce environment[6].

Due to the lack of studies in this field this study proposes a new procedure for discovering automatically the users' interests basing on their browsing behavior. This behavior is defined as actions performed by users while browsing a website, such as moving a mouse, scrolling, clicking, selecting a text, bookmarking and printing pages. These actions happen when users are interacting with the website through a web browser interface. Logging user's behavior and user preferences on the client side is a challenging task. For the purpose of this study the extension (Addon) called ECPM (E-commerce Customers' Preference Monitor) for Mozilla FireFox was built.

The rest of this paper is organized as follows; Section 1 presents related work on implicit indicators; Section 2 provides the description of the proposed framework for monitoring the user behavior and learning user preferences from these data.

1. Related Work

Implicit feedback techniques have been used for retrieving, filtering and recommending a variety of items: web sites, articles (both academic and journal), email messages, music, movies and products from e-commerce stores. Researchers have used different source data and different construction of their research environment. On a client-side technique researchers used tools for monitoring user behavior such as a browser specially created for the sake of the study, a modified browser (e.g. modified Internet Explorer), a script in Javascript language attached to a browser. Besides monitoring the user behavior in most studies participants were asked to explicitly indicate their interest in visited resources.

Monitored behavior can be classified into six categories such as: Examination, Retain, Reference, Annotate, Create[7]. For instance, reading is considered as an action that allows one to Examine an object. It is worth noting that confidence in the behaviors that is available for inference varies according to category of the observed behavior. For instance behaviors from the Examination category is a weaker evidence of user interests than behaviors from the Annotate or Create category. However most often observed actions belong to the Examination category,

which expresses the weakest evidence of the user's interest. On the other end of scale there are behaviors such as creating, printing, saving which occur less frequently.

According to a two dimensional classification proposed by Kim and Orad every action can also be to classified to the minimum possible scope of item being acted upon. For instance, Bookmarking can be performed on Object, while Reading can be observed on Segment of Object. This classification highlights that behaviors observed at a larger scale are less precise. For instance, obtaining implicit feedback about a class of objects will presumably provide less precise information about the users' interests than obtaining implicit feedback about an object or a segment of objects.

Most widely explored behaviors for implicit feedback are a document (web page) selection and the time spent on a web pages. The finding that users spend more time on documents that they find relevant has been replicated in a large number of studies[8]. However, some researchers indicate that time may not be a reliable interest indicator[9]. It can be caused by the absence of user or by performing another task while a website is open (i.e. checking e-mail, watching a movie, using a communicator, exploring another website). Therefore in the proposed framework the time spent by the user on a web page is divided into three types depending on the total time the page is loaded (Complete Duration), if the browser tab is active (Active Tab Duration), and if the user is actively interacting with the webpage (User Activity Duration). Another mostly explored implicit feedback indicators consist of a mouse-move distance, a scroll distance, a key input.

Claypool et. al[10] created a CuriousBrowser which was a web browser that recorded implicit actions and explicit ratings by users. The browser was used to record mouse clicks, mouse movement, scrolling and elapsed time. The results of these studies indicate that the time spent on a page, the amount of scrolling on a page, and the combination of time and scrolling has a strong correlation with an explicit interest.

Velayathan and Yamada[11] developed a logging tool called the Ginis site logging tool to monitor and log the user browsing behavior. They performed a user experiment using ten participants to observe the browsing behavior, and evaluated the behaviors by performing classification learning by means of the C4.5 algorithm. This tool is capable of monitoring a large number of user behaviors. They successfully confirmed that all the rules generated by the C4.5 classification learning algorithm have 60% or higher fitness, where the error for all the tests was lower than 40%.

A complete review of research into implicit feedback techniques is provided by Diane Kelly[12].

2. Framework

In order to evaluate the application of implicit feedback to learn the user preferences on e-commerce websites, the procedure with ECBE extension for FireFox has been proposed. This procedure consists of two stages: the data collection stage and the learning stage. The main goal of the first stage is to collect data concerning the user behavior on the real e-commerce websites. At the second stage the collected data are analyzed using machine learning algorithms and the rough sets theory, in order to search for patterns indicating the users' interest.

The first stage of the research involves a study with participants, as the collection of behavior data is performed via ECBE extension. Because of this every participant of the study is required to use the FireFox browser and to install the research extension.

During the preparation phase to this study the author decided on three requirements which should be met in order to produce valuable results. The first requirement concerns a tool used for data collection. Most previous researchers collected data by means of modified or specially build web browsers. In the author's opinion the introduction of a new browser with its new interface and functionality may influence the behavior of users and change their attitude towards the study. Because of this fact the author's goal was to use the existing browser with minor modifications which would allow proper data collection. The second requirement states that the participants should visit real e-commerce websites with established position on the market. Creating fully functional online shop for the purpose of the study is a challenging task. Moreover, the participants should be allowed to visit not only one website. Ideally the participants should be allowed to visit any available e-commerce website. The last study requirement is the ability to collect only these behavior data which could be collected by means of the techniques which do not require users to install any extra software. This was introduced in order to allow application of study results on real websites. Another part of this is the requirement is to collect explicit ratings provided by the participants.

After analyzing the available solutions the author decided that an ideal candidate which satisfies all the requirements is the FireFox (FF) browser. FF is a very popular browser which allows to install external extension called Addons to enhance its functionality. The FF extension uses mainly two languages JavaScript for and XUL for Interface. The FF extension called ECPM (E-commerce Customers Preference Monitor) was built for the purpose of the study to allow monitoring of various users behaviors by utilizing the Events DOM model.

The ECPM extension contains a Preference window where the user can deactivate the extension, provide personal data, open a page with the study description, open a poll window

and finish the study by sending results to the server. During the study the ECPM monitors the user behavior which are used to calculate 19 implicit indicators listed in Table 1. As it was mentioned before, all of these indicators can be calculated without using the extension just by inserting regular JavaScript code to the html page. This allows to utilize the results of the study in commercial applications.

Table 1. An assessment of didactic hours dedicated to the SMS-B system in six Faculties of Warsaw University of Life Sciences

Implicit indicator	Indicator description
Search result visit	Boolean value indicating whether the source of page visit has been a website search result page.
Complete duration	Time between page load and page unload
Active tab duration	Amount of time (ms) while tab containing particular page is being active
User activity duration	Amount of time (ms) while user is actively interacting with page thus generates keyboard and mouse events
Description area mouse duration	Amount of time (ms) while mouse pointer is positioned over description of the product
Picture area mouse duration	Amount of time (ms) while mouse pointer is positioned over pictures of the product
Review area mouse duration	Amount of time (ms) while mouse pointer is positioned over users' review of the product
Recommendation area pointer duration	Amount of time (ms) while mouse pointer is positioned over other recommended products
Mouse movement distance	Total distance (in pixels) of mouse pointer's movement
Horizontal scroll distance	Total distance (in pixels) of horizontal scroll of page content
Vertical scroll distance	Total distance (in pixels) of vertical scroll of page content
Total mouse click	Total number of mouse clicks regardless which mouse key has been pressed
Left MB click	Total number of left mouse button clicks
Right MB click	Total number of right mouse button clicks
Middle MB click	Total number of middle mouse button clicks
Copy/Cut action	Total number of copy/cut actions performed via keyboard shortcut
Save	Total number of save actions performed via keyboard shortcut
Print	Total number of print action performed via keyboard shortcut
Bookmark	Total number of bookmarking actions performed via keyboard shortcut
Zoom	Total number of zooming actions performed via keyboard shortcut
Find	Total number of found actions performed via keyboard shortcut
Content select action	Total number of page content selection actions
Text select size	Total number of selected chars
Keydown action	Total number of key pressings

Source: own study.

Among the indicators which were previously used in other experiments this study introduces five new indicators, such as: User Activity Duration, Description Area Mouse Duration, Picture Area Mouse Duration, Review Area Mouse Duration and Recommendation

Area Mouse Duration. The User Activity Duration is calculated as a sum of the time period of two seconds while the user has actively interacted with a website. Very often users open many pages in different tabs or leave their computer while the website is open. The User Activity Duration is supposed to express the period of time while the user is actively using the website. Nowadays, almost every e-commerce product page contains following sections: a product description, product pictures, a review and users' opinions about the product as well as a recommendations area where recommended products are displayed. The author hypothesizes that reading these sections may indicate the user's interest. Therefore this study proposes four indicators calculated as the time when a mouse pointer is placed within these areas. Because of the fact that the HTML structure of every page is constructed differently, the method for calculating these indicators must be computed differently. The ECPM extension was prepared for proper calculations in five Polish e-commerce shops selected for the study. The structure of every website is stored within the extension. The ECPM can be further extended to allow for complete monitoring of other websites.

Beside non-relative indicators which were calculated directly from the user behaviors, indicators relative to page measurements were also constructed. In order to calculate these indicators the following measures are calculated for every website: total page characters length, total recommendation area characters length, total recommendation area characters length, page height, number of product pictures, visible page height. By using behavior data and page measure data the relative indicators are calculated.

When user is leaving the product page the Interest evaluation window is displayed (Figure 1). This window contains two questions concerning the product, the page of which is being left. Originally this study is designed for Polish language speaking participants. First question "How much does this product interest you?" allows the user to express his/her explicit interest in this product. This interest can be rated on a five point scale from 1–5 with 5 as *very much* and 1 as *not at all*. Another question "Have you known this product before?" checks if the user is familiar with the viewed product. After the form has been filled in, the answers are stored. If it happens that the user returns to this product page again during the survey, all the previously marked answers will be displayed.

The structure and functionality of the EIP portals based on standardized business project methodologies: B2B (Business-to-Business), B2C (Business-to-Consumer), B2E (Business-to-Employee), B2G (Business-to-Government). The EIP portals offer a single point of entry and user authentication to information systems and data storages in institutions basing on WWW pages technologies.

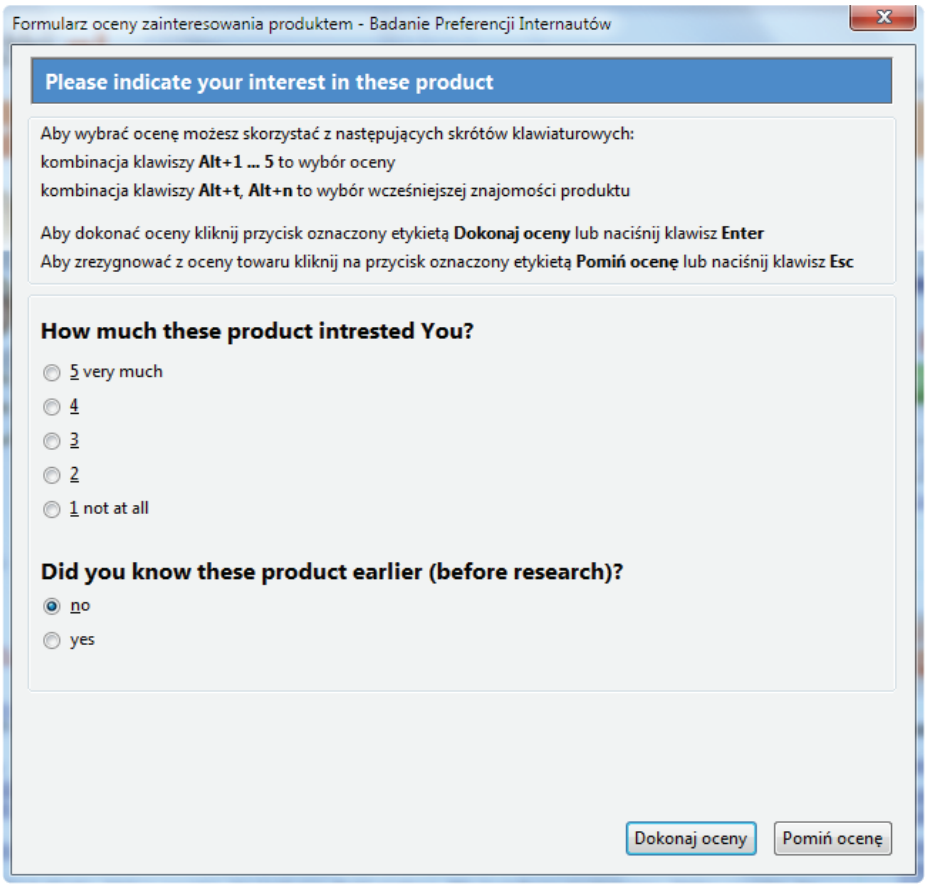

Fig. 1. Explicit Interest evaluation form.

Source: own study.

Conclusion

The preliminary study method was tested on a group of 17 students of Computer Science. The participants were not aware of the purpose of the study, in order to avoid any influence on their behavior. The participants were given a task to look for interesting products in five popular Polish e-commerce shops: komputronik.pl, agito.pl, electro.pl, merlin.pl, morele.net. During the study the participants evaluated 461 product pages. After preliminary cleaning the data was divided into two sets: learning 70% and evaluation 30%. The initial classification tree model was built on these data showing 46.3% of misclassification rate. The most important for e-commerce shops is information about products most liked by customers, then the interest ratings from 4

to 5 were joined together into the value 1, which meant strong interest, and ratings from 1 to 3 were joined into the value 0, which meant low or no interest. The classification tree built on the basis of these data improved the classification to 72.3%.

In the future the author is planning to apply other machine learning algorithms such as a rough set theory or neural networks. Classification will be performed on different sets of data containing: all samples, samples for every user and samples for every website in order to test the hypothesis about different patterns of behavior among different people and among different websites.

Notes

[1] Nichols. D.M. (1997)..

[2] Middleton, S.E., Shadbolt, N.R., De Roure, D.C. (2003); Nichols (1997).

[3] Morita, M., Shinoda, Y. (1994); Kim, K., Carroll, J.M., Rosson, M. (2002); Konstan, J., Miller, B., Maltz, M., Herlocker, J., Gordon, L., Riedl, J (1997).

[4] Case, D. (2012).

[5] Oard, D.W., Kim, J. (2001); Kelly, D., Belkin, N.J. (2001); Velayathan, G., Yamada, S. (2005).

[6] Kim, Y.S., Yum, B.J., Song, J., Kim, S.M. (2005); Peška, L., Vojtáš, P. (2012); Zhang, Z., Qian, S. (2012).

[7] Oard, D.W., Kim, J. (2001).

[8] Ibidem; Cooper, M.D., Chen, H.-M. (2001); Miller, B.N., Riedl, J.T., Konstan, J.A. (2003).

[9] Seo, Y.W., Zhang, B.T. (2000); Jung, K. (2001).

[10] Kelly, D., Belkin, N.J. (2001).

[11] Claypool, M., Le, P., Wased, M., Brown, D. (2001).

[12] Velayathan, G., Yamada, S. (2005).

References

Case, D. (2012). *Looking for Information: A Survey of Research on Information Seeking, Needs and Behavior (Library and Information Science)*. Emerald Group Publishing.

Claypool, M., Le, P., Wased, M., Brown, D. (2001). *Implicit interest indicators*. In Proc. 6th international conference on Intelligent User Interfaces.

Cooper, M.D., Chen, H.-M. (2001). Predicting the Relevance of a Library Catalog Search. *Journal of the American Society for Information Science*, 52 (10).

Jung, K. (2001). *Modeling web user interest with implicit indicators*. Master Thesis, Florida Institute of Technology, USA.

Kelly, D. (2005), *Implicit Feedback: Using Behavior to Infer Relevance*. New directions in cognitive information retrieval, The Information Retrieval Series, Volume 19, Section IV.

Kelly, D., Belkin, N.J. (2001). *Reading time, scrolling and interaction: exploring implicit sources of user preferences for relevance feedback*. In SIGIR '01.

Kim, K., Carroll, J. M., Rosson, M. (2002). *An Empirical Study of Web Personalization Assistants: Supporting End-Users in Web Information Systems*. In Proceedings of the IEEE 2002 Symposia on Human Centric Computing Languages and Environments. Arlington, USA.

Kim, Y.S., Yum, B.J., Song, J., Kim, S.M. (2005). *Development of a recommender system based on navigational and behavioral patterns of customers in e-commerce sites*. Expert Systems with Applications Volume 28, Issue 2.

Konstan, J., Miller, B., Maltz, M., Herlocker, J., Gordon, L., Riedl, J (1997). *GroupLens: Applying Collaborative Filtering to Usenet News*. Communications of the ACM, 40(3).

Middleton, S.E., Shadbolt, N.R., De Roure, D.C. (2003). *Capturing Interest through Inference and Visualization: Ontological User Profiling in Recommender Systems*. In Proceedings of the Second Annual Conference on Knowledge Capture 2003.

Miller, B.N., Riedl, J.T., Konstan, J.A. (2003). *GroupLens for Usenet: Experiences in Applying Collaborative Filtering to a Social Information System*. In: C. Lueg and D. Fisher [Ed.], *From Usenet to CoWebs: Interacting With Social Information Spaces*. London: Springer Press.

Morita, M., Shinoda, Y. (1994). *Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval*. In Proceedings of ACM Conference on Research and Development in Information Retrieval (SIGIR '94), Dublin, Ireland.

Nichols, D.M. (1997). *Implicit Ratings and Filtering*. In Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering (pp. 31–36), Hungary.

Oard, D.W., Kim, J. (2001). *Modeling Information Content Using Observable Behavior*. In Proceedings of the 64th Annual Meeting of the American Society for Information Science and Technology (ASIST '01), USA.

Peška, L., Vojtáš, P. (2012). *Estimating importance of implicit factors in e-commerce recommender systems*. ACM WIMS '12.

Seo, Y.W., Zhang, B.T. (2000). *A Reinforcement Learning Agent for Personalized Information Filtering*. In Proceedings of the 5th International Conference on Intelligent User Interfaces, USA.

Velayathan, G., Yamada, S. (2005). Behavior Based Web Page Evaluation. *Journal of Web Engineering*, 1, 1.

Zhang, Z., Qian, S. (2012). *The Research of E-commerce Recommendation System Based on Collaborative Filtering Technology*. Advances in Intelligent and Soft Computing Volume 168.