Journal of The Korea Society of Computer and Information
Vol. 20, No. 3, March 2015
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2015.20.3.047

SVM과 로짓회귀분석을 이용한 흥미있는 웹페이지 예측

전 도 홍*, 김 형 래**

Predicting Interesting Web Pages by SVM and Logit-regression

Dohong Jeon*, Hyoungrae Kim**

요 약

흥미 있는 웹페이지의 자동화된 탐색은 다양한 응용 분야에 활용될 수 있다. 웹페이지에 대한 사용자의 흥미는 판단하는 것은 사용자의 행동을 관찰함으로 자동화가 가능하다. 흥미 있는 웹페이지를 구분하는 작업은 판별 문제에 속하며, 우리는 실증을 위해 화이트 박스의 학습 방법(로짓회귀분석, 지지기반학습)을 선택한다. 실험 결과는 다음을 나타내었다. (1) 고정효과 로짓회귀분석, polynomial 과 radial 커널을 이용한 고정효과 지지기반학습은 선형 커널보다 높은 성능을 보였다. (2) 개인화가 모델 성능을 향상시킴에 있어 주요한 이슈이다. (3) 사용자에게 웹페이지에 대항 흥미를 물을 때, 구간은 단순히 예/아니 도 충분할 수 있다. (4) 웹페이지에 머문 기간이 매초 증가할 때마다 성공확률은 1.004배 증가하며, 하지만 스크롤바 클릭 수 (p=0.56) 와 마우스 클릭 수 (p=0.36) 지표는 흥미와 통계적으로 유의한 관계를 가지지 않았다.

▸ Keywords : 기계학습; 자동화 된 지표; 웹 페이지; 흥미

Abstract

Automated detection of interesting web pages could be used in many different application domains. Determining a user's interesting web pages can be performed implicitly by observing the user's behavior. The task of distinguishing interesting web pages belongs to a classification problem, and we choose white box learning methods (fixed effect logit regression and support vector machine) to test empirically. The result indicated that (1) fixed effect logit regression, fixed effect SVMs with both polynomial and radial basis kernels showed higher performance than the linear kernel model, (2) a personalization is a critical issue for improving the performance of a model, (3) when asking a user explicit grading of web pages, the

•제1저자 : 전도홍 •교신저자 : 김형래
•투고일 : 2015. 1. 7, 심사일 : 2015. 2. 2, 게재확정일 : 2015. 2. 24.
* 가톨릭관동대학교 컴퓨터학과(Dept. of Computer Science, Catholic Kwandong University)
** 한국고용정보원(Korea Employment Information Service)

www.kci.go.kr

scale could be as simple as yes/no answer, (4) every second the duration in a web page increases, the ratio of the probability to be interesting increased 1.004 times, but the number of scrollbar clicks (p=0.56) and the number of mouse clicks (p=0.36) did not have statistically significant relations with the interest.

▸ Keywords : machine learning; Implicit indicator; Web pages; Interest

I. Introduction

Determining a user's interesting web pages can be performed explicitly by asking the user, or implicitly by observing the user's behavior [1]. Implicit indicators are usually less accurate than explicit indicators [2]. However, implicit indicators do not require any extra time or effort from the user and can adapt to changes in the user's interests over time. To implicitly measure user interest we need to identify reliable implicit indicators. One of the major user interest indicators identified by researchers is duration, or the time spent on a web page [3-9]. The previous researches mainly focused on identifying reliable indicators. However this research more focus on learning with the indicators.

Automated detection of interesting web pages could be used in many different application domains [10,11]. The problem is to find a proper learning method for interesting web pages. The inputs for the learning method are web log-files and interest scores of web pages provided by a user. The log-files record users' behavior while they are reading web pages. The learning method will belong to a supervised learning.

First we select implicit indicators for identifying interesting web pages. In order to learn which machine learning method is proper for the purpose of this research. The task of distinguishing interesting web pages belongs to a classification problem, and we choose some learning methods to test empirically. Since white box learning methods help us understand which indicators are more informative and how the classification methods works [12], we prefer white box learning. The black box learning, however, shows relatively higher performance but is hard to understand. And the behaviors are different depending on each user, we control these personal noises by adopting fixed effect model.

The main contributions of this research are:

- when compared logit regression and SVM models, logit showed the highest log likelihood (-123.1), but SVM with polynomial and radial basis kernel implied to have a higher performance depending on application domain over a ROC curve demonstration;

- as all fixed effect models performed higher than the mixed effect models, this indicated that a personalization is a critical issue for improving the performance of a model;

- as the weight information does not give significant improvement in performance, this implied that when asking a user explicit grading of web pages, the scale could be as simple as yes/no answer;

- detail analysis of logit regression showed that every second the duration increases, the ratio of the probability to be interesting increased 1.004 times, but the NumberOfScrollbarClick (p=0.56) and NumberOfMouseClick (p=0.36) did not have statistically significant relation with the interest.

The rest of this paper is as follows: Section II discusses related work in implicit indicators for

www.kci.go.kr

interesting web pages; Section III introduces implicit indicators; Section IV details our machine learning approach; Section V describes our experiments; Section VI analyzes the results from the experiments; Section VII summarizes our findings and suggests possible future work.

II. Related works

The previous research [13] mainly focus on identifying significant indicators. Jung [4] developed Kixbrowser, a custom web browser that recorded users' explicit rating for web pages and their various actions. His results indicate that the number of mouse clicks is the most accurate indicator for predicting a user's interest level. Goecks and Shavlik [14] proposed an approach for an intelligent web browser that is able to learn a user's interest without the need for explicitly rating pages. They measured mouse movement and scrolling activity in addition to user browsing activity (e.g., navigation history). The indicator of hyperlink clicked showed the lowest RMS errors. CuriousBrowser [5] is a web browser that recorded the actions (implicit ratings) and explicit ratings of users. This browser was used to record mouse clicks, mouse movement, scrolling and elapsed time. The results 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 explicit interest. Sometimes results from different researchers showed some inconsistency. The mouse click is a good indicator, but Claypool et al. [5] did not in Jung's [4]. The scrollbar movement also showed inconsistency depending on the researchers (Jung, 2001; Claypool et al. [5] Powerize [8] reported a way to implement the implicit feedback technique of user modelling for Powerize. They also found that observing the printing of web pages along with reading time could increase the prediction rate for detecting relevant documents.

Another type of analysis is to use the rank of the search results instead of an explicit rating of interests. Granka et al. [3] measured eye-tracking to determine how the displayed web pages are actually viewed. Their experimental environment was restricted to a search results. They analysed the relation between the rank in the search result and the interests of the web pages.

Our analysis focuses more on predicting the probability of interest to users. This task is related to both a machine learning and a statistical analysis.

III. Implicit Indicators

This section describes indicators of duration, as well as other user interest indicators that will be examined.

3.1. Duration

A user may tend to spend more time on pages that he or she finds interesting, so we record the duration spent on a web page. The complete duration is defined as the time interval between the time a user opens and leaves a web page. Some web pages contain many images that delay the downloading time, so we start measuring the duration after the entire page is loaded. Thus, the complete duration won't be affected by the connection speed, the amount of Internet traffic, or the CPU speed. The complete duration for a web page can be calculated by subtracting the time of finishing downloading the current web page from the time of leaving the web page. The complete duration is different from the duration used by Jung [4]. His duration includes the downloading time of a web page.

3.2. Distance of Mouse Movement

Many people move their mouse while reading the contents of a web page. Mouse movement can occur while looking at an interesting image, or when

www.kci.go.kr

pointing at interesting objects. We hypothesize that the more distance a mouse moves, the more a user be interested in the web page. The distance of mouse movement is detected by its x and y coordinates on a monitor every 100 milliseconds. The formula is

$$mouse\_movement(pixels) = \sum_{i=1}^{t-1} Dist(P(t_i) - P(t_{i-1})) \quad (1)$$

where $P(t_i)$ is a mouse location with x and y coordinates at time t_i, and the Dist function is a Euclidean distance.

3.3. Number of Mouse Clicks

People use "click" to hyperlink to another web page. In addition, clicking can be considered as a habitual behavior. Clicking can be a way of expressing our emotions such as if some people are happy to find a product that they were looking for (e.g., book), then they can click the object several times repeatedly. This indicator was examined in Kixbrowser [4], Curious browser [5], Goeck's browser [14], and Letizia [7]. We use the hypothesis that the greater the number of mouse clicks on a web page is, the more a user is interested in it.

3.4. Distance of Scrollbar Movement

A user can also scroll a web page up and down by dragging a scrollbar. Those dragging events can occur several times while a user is reading a web page. The distance of a scrollbar movement for a web page is the sum of all distances of scrollbar movement for all occasions.

$$scrollbar\_movement(pixels) = \sum_{j}^{E} \sum_{i=1}^{E(j)-1} |P(t_i) - P(t_{i-1})| \quad (2)$$

where E is the number of times the scrollbar is pressed, time E(j) is the duration that the scrollbar is dragged in a single dragging event.

3.5. Number of Scrollbar Clicks

The length of many web pages is longer than the height of a monitor. If a user finds a web page interesting, he or she may read further down the web page. A user can scroll down a web page either by clicking or by dragging the scrollbar. As a user scrolls a web page up and down by clicking, the number of scrollbar clicks increases. Jung [4], Goecks et al. [14], and Claypool et al. [5] measured this event and reported that it is a good indicator.

3.6. Number of Key UP and Down

When scrolling a web page, some people use the "up" and "down" keys instead of the scrollbar. The hypothesis is that the greater the number of key up and down presses, the more a user is interested in the web page. This event is measured by increasing the count every time a user strikes up or down keys. Curious browser [5] and Jung [4] measured keyboard activities.

3.7. Size of Highlighting Text

While reading a web page, if a user copies some contents of the web page it probably means that the user is interested in the web page. Furthermore, a user can also habitually highlight portions of the page that they are interested in, which is a sign that the user is interested in the page. We assume that the more a user highlights in a web page, the more a user is interested in that web page. A user can highlight several different sentences in a web page for several different occasions. We sum all highlighted contents at the end. We assumed a character is 5 pixels, each line has 80 characters, and distance between two lines is 20 pixels on average. The formula is

$$highlighting\_text = \sum_{j}^{E} DistY_j / 20 \times 80 + DistX_j / 5 \quad (3)$$

www.kci.go.kr

where E is the number of occasions when highlighting occurs, DistY is the vertical distance between two points, and DistX is the horizontal distance between two points.

3.8. Other Indicators

We also measure bookmark, save, and print. We assume bookmarked web pages are interesting to a user [15,16]. Users save important/interesting web pages in their personal storage by using the "Save As" command. This also implies that those saved web pages are interesting to users [7]. The printed web pages are likely to be interesting to users [8].

IV. Learning Interest Indicators

The purpose of learning is to predict the interests of a user towards a new web pages by the indicators of user behaviors. This task of distinguishing interesting web pages belongs to a classification problem [18]. Among several classification methods we prefer white box learning. White box learning help us understand which indicators are more informative and how the classification methods works [12]. These advantages of white box methods over black box ones makes our research results more applicable to other different research areas such as identifying interesting items to users. The results saying which indicators are how much important towards others they will provide incites to developers when choosing indicators to detect user behaviors more efficiently.

Among several white box machine learning algorithms we choose logistic regression (logit-regression) and support vector machines (SVM). The reason of choosing these algorithms are due to the characteristics of the data we are analysing. This data is about human behavior, in which the distributions have meaning. Another characteristics of this is that the behavior towards interesting web pages may differ over different users. In order to control the effect caused by different users, we also try a fixed effect logit regression model [17].

$$y_i = \frac{1}{1+e^{\beta_0 + \sum_{k=1}^{p}\beta_k I_k + \beta_k U_k + \epsilon_i}},\ \epsilon_i \sim N(0,\sigma^2) \quad (4)$$

where I denotes indicators and U implies users.

We compare this result with fixed effect support vector machine. SUM uses only support vectors and be infamous for its high accuracy [2]. The fixed effect support vector machine controls the effect by different users. Many research applied kernel trick to improve the performance of SVM [2]. We apply the most well known three kernels : linear, polynomial, and radial basis. The type of SVM is C-classification, since we adjusted the dependent variable to be categorical one.

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j K(x_i, x') - \sum_{i=1}^{m}\alpha_i$$
$$\text{s.t.} \quad \sum_{i=1}^{m} y_i \alpha_i = 0,\ D \geq \alpha_i \geq 0 \quad i = 1,..,m \quad (5)$$

- linear kernel: $x_i \circ x'$ (6)
- polynomial kernel: $(\lambda x_i \circ x' + \beta_0)^3$ (7)
- radial basis kernel: $e^{-\lambda \|x_i - x'\|^2}$ (8)

where, For linear kernel, the cost value is 1 which is the default value. For polynomial kernel, the degree is 3, gamma(λ) is 1 divided by the number of attributes. For radial basis kernel we user the same gamma value as polynomial kernel.

The input values are the indicators: complete duration (Complete), distance of mouse movement (MousMove), number of mouse clicks (MousClk#), distance of scrollbar movement (ScrolMov), number of scrollbar clicks (ScrolCk#), number of key up and down (KeyUpDn#), and size of highlighting text (Highligh), bookmarked, saved, printed. The output

www.kci.go.kr

value will be the interest of a user to a web page.

V. Experiments

For our experiments, we built a web browser that can record the indicators described above from user's behavior. Data sets were collected from 12 different users. Each user was asked to spend a total of 2 hours at the computer. All volunteers were encouraged to behave as usual. To get a variety of behaviors, we asked the volunteers to divide their activities into multiple sessions, each of which does not exceed 1 hour.

For web pages that a user visited more than once, the score might be the same, but all other information (the durations or number of mouse clicks etc.) may be different. We use the duration of the view where the user stayed for the longest period of time, because users do not tend to read the web page again if they know about a web page before [9]. On average, users had 182 visits in the "visits with maximum duration" data set.

Every time a user leaved a web page, the web browser asked the user how much they are interested in the web page - there were 5 scales between "not interested" (1) and "very interested" (5). The interests were subjective to each user. The system had a "rescore" button to allow changing the score marked in the previous visit. The browser was written in Visual Studio .NET and ran on a Pentium 4 CPU. The Operating System was Windows XP.

In order to measure the performance of learning methods, 10 fold evaluation method is used. We measured how accurate an indicator could predict a user's interest. We used ROC (receiver operating characteristic curve) and log likelihood. ROC plot both the true positive rate and the false positive rate [11]. This curve represents the trade-off between sensitivity (true positive or recall rate) and specificity (true negative rate). The specificity is complementary to the false positive rate, so 1-specificity becomes false positive rate. ROC is measured by AUC (area under the curve). The higher AUC, the better ROC. Log likelihood is another important evaluation [11]. The log likelihood is the logarithm of the product of the probability the method predicted to each class. The log likelihood value is always negative, and is better as we get the value closer to 0.

VI. Results and Analysis

This section analyzes the data collected from the users who participated in our experiment.

6.1. Performance measure by ROC and log likelihood

We compared four models: logit regression and SVM with three different kernels. In order to see the difference between mixed model and fixed model, we presented the whole results in Table 1. The eight different results (2 different effects and 4 models) can be easily compared in the list. Fixed effect logit regression yield the highest AUC value of 0.726. It's log likelihood is also the closest to 0 (-123.1). It was difficult to determine the second highest model. Because the fixed effect SVM with polynomial kernel had the second highest AUC (0.678), but the fixed effect SVM with radial basis kernel had the second closet log likelihood value(-126.6) to 0. Even though the difference of AUC and log likelihood among different model, the result indicated that the fixed effect logit regression had the highest performance.

The fixed effect model showed higher performance than mixed effect model in all four models. The fixed effect model controled the effect of the user differences. Instead of mixing all users' data sets together, each individual data set was analysed separately so that we could clearly observe whether some indicator predicted certain individual's interests more accurately than other indicators. This result implied that the learning model of implicit indicator had to be personalized for each user.

www.kci.go.kr

We presented the ROCs of fixed effect logit regression model and other 3 methods in Figure 1. One can argue that fixed effect SVM with polynomial kernel and fixed effect SVM with radial basis kernel produce higher True positive rate over lower false positive rate relatively. One can also insist that the SVM models which are closer to the left top corner will give higher performance. We would like to leave this decision to the reader, since it may depend on which point you may want to choose.

Table 1. Comparing learning methods by AUC and log likelihood

Category	Learning method	AUC	log likelihood
Mixed effect	logit regression	0.616	-131.6
	SVM with linear Kernel	0.553	-132.7
	SVM with polynomial Kernel	0.621	-132.8
	SVM with radial Kernel	0.565	-130.2
Fixed effect	logit regression	0.726	-123.1
	SVM with linear Kernel	0.664	-126.7
	SVM with polynomial Kernel	0.678	-130.0
	SVM with radial Kernel	0.661	-126.6

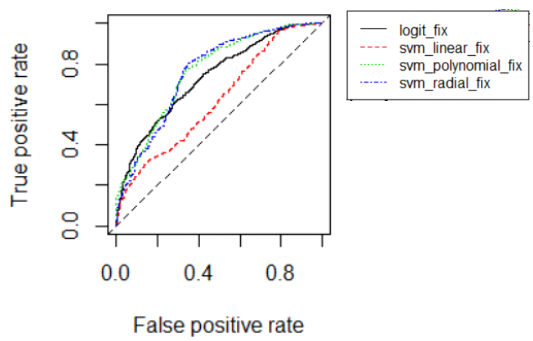

Figure 1. ROC curve of Logit regression with fixed effects

6.2. No-weighted vs. weighted dataset

The fixed effect logit regression could be weighted. The weights were derived from the scale of the answer. Both scales of "not interested" and "very interested" was weighted as 3 and the "interested" in the middle was weighted as 1. We presented the both results from two models (not-weighted and weighted) for comparison purposes in Table 2 and Table 3. The Chi-square test told us that both model were statistically significant, since the p-value is much lower than 0.05. The model not-weight had log likelihood value of -1051.123 which is closer to 0 than the log likelihood value of -2128.818 from the weighted model. The smaller AIC (Akaike Information Criterion) value is the better fit the model is [17]. The model not-weight had smaller AIC value of 2146.2. the not-weighted model showed higher best fit in mixed effect models as well yielding the Chi-square value of -1161.701 and the AIC value of 2345.4. This result indicated that when you measure the intense of the interest from a user Yes/No is good enough.

Table 2. Logit-regression analysis with mixed effect

Variables	Not-weighted				Weighted			
	coef. (standardized)	p<		odds ratio	coef. (standardized)	p<		odds ratio
(Intercept)	0.668	0.00	***	1.626	0.362	0.00	***	1.175
Duration	0.152	0.01	*	1.002	0.203	0.00	***	1.003
DistanceOfMouseMovement	0.169	0.00	**	1.000	0.170	0.00	***	1.000
NumberOfScrollbarClick	-0.109	0.23		0.980	-0.127	0.03	*	0.976
DistanceOfScrollbarMovement	-0.014	0.73		1.000	-0.018	0.53		1.000
NumberOfKeyUpDown	-0.081	0.01	*	0.969	-0.113	0.00	***	0.957
NumberOfMouseClick	0.052	0.61		1.008	0.063	0.37		1.010
SizeOfHighlightingText	-0.045	0.24		1.000	-0.045	0.08	.	1.000
Bookmarked	0.549	0.00	**	1.731	0.791	0.00	***	2.205
Saved	0.254	0.38		1.290	0.415	0.04	*	1.515
Printed	1.055	0.00	**	2.871	1.321	0.00	***	3.747
Chisq test	1.000767e-14				1.289559e-61			
log likelihood	-1161.701 (df=11)				-2389.486 (df=11)			
AIC:	2345.400				4801.000			

* Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3. Logit-regression analysis with fixed effect

Variables	Not-weighted				Weighted			
	coef. (standardized)	p<		odds ratio	coef. (standardized)	p<		odds ratio
(Intercept)	0.805	0.00	***	1.957	0.638	0.00	***	1.618
Duration	0.306	0.00	***	1.004	0.320	0.00	***	1.004
DistanceOfMouseMovement	0.079	0.20		1.000	0.116	0.00	**	1.000
NumberOfScrollbarClick	-0.039	0.67		0.993	-0.037	0.56		0.993
DistanceOfScrollbarMovement	-0.052	0.22		1.000	-0.060	0.03	*	1.000
NumberOfKeyUpDown	0.068	0.07	.	1.027	0.068	0.01	*	1.027
NumberOfMouseClick	-0.054	0.61		0.991	-0.067	0.36		0.989
SizeOfHighlightingText	-0.050	0.22		1.000	-0.055	0.05	*	1.000
Bookmarked	0.733	0.00	***	2.081	0.959	0.00	***	2.610
Saved	0.301	0.32		1.351	0.487	0.02	*	1.627
Printed	1.124	0.00	**	3.076	1.447	0.00	***	4.249
USERs	... omitted				... omitted			
Chisq test	3.228126e-53				2.201114e-163			
log likelihood	-1051.123 (df=22)				-2128.818 (df=22)			
AIC:	2146.2				4301.6			

* Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

www.kci.go.kr

6.3. Significant Indicators in a fixed effect logit regression

It is worth to measure how much effect each indicator has in the fixed effect logit regression. In order to evaluate each indicator to see which one is more predictable, we presented coefficients (log odds ratio), p-value, and odds ratio. The coefficients were scaled to see which indicator is stronger than others, but the odds ratio is not scaled to measure the real effect. As we look at the not-weighted model in Table 3, the indicators of print, bookmark, duration had coefficient values of 1.124, 0.733, and 0.306 respectively. The odds ratio predicts the ratio of the probability of an indicator to be interested to a user. The odds ratio of these indicators implied that when a user printed a web pages the probability of being interested increased 3.076 times; when saved a web page, the ratio of the probability increased 2.081 times; everytime the duration increases a second, the ratio of the probability increased 1.004 times.

But if we use the weight we could use more number of implicit indicators. If we order the statistically significant indicators by the strength assuming the indicators were weighted, they could be listed as printed (1.447), bookmarked (0.959), saved (0.487), duration (0.32), DistanceOfMouseMovement (0.116) etc. When a user printed a web pages, the ratio of the probability that the web page become interested to a user increased more than 4.2 times than the un-printed web pages. During the stay in a web page every second increased the probability to be interested 1.0004 times.

However, the NumberOfScrollbarClick ($p=0.56$) and NumberOfMouseClick ($p=0.36$) were never statistically significant in four models. A user click the scroll bar just to read more of the web pages. This does not mean that s/he is interested in the web pages. We also assume that the mouse click is just a habitual action. This does not express an emotional happiness for finding any interesting web pages. Jung [4], Goecks and Shavlik [14], and Claypool et al. [5] reported this NumberOfScrollbarClick is a good indicator. We controlled the other effect and only measured purely by this event. This is why our result is different from previous analysis.

VII. Summary

This paper studies several implicit indicators that can be used to determine a user's interest in a web page. All indicators examined were duration, distance of mouse movement, number of mouse clicks, distance of scrollbar movement, number of scrollbar clicks, number of key up and down, size of highlighting text, bookmarked, saved, and printed.

We evaluated how accurately a model can predict users' interests by ROC and log likelihood. Among different machine learning methods, white box methods are chosen: logit regression and SVM. We compared four different methods: logit regression, SVM with linear kernel, SVM with polynomial kernel, and SVM with radial basis kernel. Each method is designed as mixed effect model and fixed effect model. The dataset can be divided into not-weighted and weighted. We used two data sets: not-weighted and weighted.

The results of AUC and log likelihood indicated that among different models the fixed effect logit regression showed the highest performance. The fixed effect model showed higher performance than mixed effect model in all four models. The fixed effect model controled the effect of the user differences. This result implied that the learning model of implicit indicator had to be personalized for each user.

The fixed effect logit regression could be weighted. The Chi-square test told us that both model were statistically significant, since the p-value is much lower than 0.05. The model not-weight had log likelihood value of -1051.123 which is closer to 0 than the log likelihood value of -2128.818 from the

www.kci.go.kr

weighted model. The smaller AIC value is the better fit the model is [17]. The model not-weight had smaller AIC value of 2146.2. This result indicated that when you measure the intense of the interest from a user Yes/No is good enough.

It is worth to measure how much effect each indicator has in the fixed effect logit regression. In order to evaluate each indicator to see which one is more predictable, we presented coefficients (log odds ratio), p-value, and odds ratio. As we look at the not-weighted model in Table 3, the indicators of print, bookmark, duration had coefficient values of 1.124, 0.733, and 0.306 respectively. The odds ratio predicts the ratio of the probability of an indicator to be interested to a user.

The odds ratio of these indicators implied that when a user printed a web pages the ratio of the probability of being interested increased 3.076 times; when saved a web page, the ratio of the probability increased 2.081 times; everytime the duration increases a second, the ratio of the probability increased 1.004 times. However, the NumberOfScrollbarClick ($p=0.56$) and NumberOfMouseClick ($p=0.36$) were never statistically significant in any model.

The limitation of this research is the low sensitivity compared to the specificity. It may due to the ratively smaller number of dataset per person. We would attempt to develop a more personalized and self-learning algorithm to improve the accuracy.

REFERENCES

[1] C. Shahabi, and F. Banaei-Kashani, "Efficient and Anonymous Web-Usage Mining for Web Personalization," INFORMS Journal on Computing-Special Issue on Data Mining, Vol. 15, No. 2, Apr. 2003.

[2] V. Kumar, "Support Vector Machines - Optimization Based Theory," Algorithms, and Extensions, Chapman & Hall/CRC Press, Dec. 2012.

[3] L. A. Granka, T. Joachims, and G. Gay, "Eye-tracking Analysis of User Behavior in WWW Search," Proc. 27th annual international conference on Research and development in information retrieval, July 2004.

[4] K. Jung, "Modeling Web User Interest with Implicit Indicators," Master Thesis, Florida Institute of Technology, Dec. 2001.

[5] M. Claypool, P. Le, M. Wased, and D. Brown, "Implicit Interest Indicators," Proc. 6th international conference on Intelligent User Interfaces, pp. 33-40, Jan. 2001.

[6] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An open Architecture for Collaborative Filtering of Netnews," In Richard K. Faruta and Christine M. Neuwirth, editors, Proc. Conference on Computer Supported Cooperative Work, ACM, pp. 175-186, Oct. 1994.

[7] H. Liberman, "Letizia: An Agent that Assists Web Browsing," Proc. 14th International Joint Conference on Artificial Intelligence, Montreal, Aug. 1995.

[8] J. Kim, D. W. Oard, and K. Romanik, "Using Implicit Feedback for User Modeling in Internet and Intranet Searching," Technical Report, College of Library and Information Services, University of Maryland, May 2001.

[9] M. Pazzani, and D. Billsus, "Adaptive Web Site Agents," Proc. 3rd International Conference, Autonomous Agents, Seattle, Washington, May 1999.

[10] S. Zahoor, M. Bedekar, P. K. Kosamkar, "User Implicit Interest Indicators learned from the Browser on the Client Side," Pro. of the 2014 International Conference on Information and Communication Technology for Competitive Strategies, Nov. 2014.

[11] N. Zumel and J. Mount, "Practical Data Science with R," Manning, pp. 101-104, Mar. 2014.

[12] T. Mitchell, "Machine Learning," McGraw-Hill, pp. 81-126 and pp. 154-199, Mar. 1997.

www.kci.go.kr

[13] H. Kim, P. K. Chan, "Implicit Indicators for Interesting Web Pages," International Conference on Web Information Systems and Technologies, Miami, Florida, USA. WEBIST press, pp. 270-277, May 2005.
[14] J. Goecks, and J. W. Shavlik, "Learning Users' Interests by Unobtrusively Observing Their Normal Behavior," Proc. ACM Intelligent User Interfaces Conference (IUI), Jan. 2000.
[15] H. Seo, J. Kim, "Development of a Robot Performance System Employing a Motion Database," Journal of Korea Society of Computer and Information, Vol. 19, No. 12. pp. 21-29, Dec. 2014.
[16] Y. S. Maarek, I. Z. B. Shaul, Automatically organizing bookmarks per contents, Computer Networks and ISDN Systems, 28 (7), 1321-1333, May 1996.
[17] D. E. Hinkle, W. Wiersma, and S. G. Jurs, "Applied Statistics for the Behavioral Sciences (4th ed.)," Boston: houghton Mifflin, Jan. 1998.
[18] G. Heo and S. Kim, "A New Clustering Method for Minimum Classification Error," Journal of The Korea Society of Computer and Information, 19(7), July 2014.

저 자 소 개

전 도 홍

1985: Olahoma City Univ. 컴퓨터과학 학사
1987: Florida Inst. of Tech. 컴퓨터과학 석사
1990: Florida Inst. of Tech. 컴퓨터과학 박사
현재: 가톨릭관동대학교 컴퓨터학과 교수
관심분야: Computer graphics, Big data, Data mining
Email : dhjeon@kd.ac.kr

김 형 래

1997: 관동대학교 컴퓨터과학 학사
2005: Florida Inst. of Tech, 컴퓨터과학 석박사
현재: KEIS, Research Fellow
관심분야: Machine learning, Robot intelligence
Email: goddoes8@gmail.com

www.kci.go.kr