Predicting User Interests Based on Their Latest Web Activities

TSUCHIYA Takeshi, HIROSE Hiroo, YAMADA Tetsuyasu
Dept. of Applied Information Engineering
Suwa University of Science
Chino city, Nagano, Japan
{ *tsuchiya, hirose, yamada* }*@rs.sus.ac.jp*

YOSHINAGA Hirokazu
Logly Inc.
Tokyo, Japan
yoshinaga@logly.co.jp

KOYANAGI Keiichi
Waseda University
Kitakyushu City, Fukuoka, Japan
keiichi.koyanagi@waseda.jp

***Abstract*—This paper discusses and proposes a method for predicting information interesting to users based on their recent online behavior. Analysis of user activities on the web aims to investigate and acquire some information about user interests through websites. Therefore, we assume that recent interests of users can be predicted by analyzing the characteristics of acquired web contents. Our proposed method identifies these user interests based on the clicked log of web advertisement by using neural networks, and makes it possible to predict information by regression to the learned user model. It means that flexible information service can be constructed using predictions based on user presence. The evaluation indicates that the method is effective and practical in comparison to the conventional model which statistically analyzes web activities.**

Keywords - user interests; content prediction; web advertisement.

I. INTRODUCTION

Online advertising, which is the essential business model, is interactive in comparison to conventional billboard advertisement and TV commercials. A representative example is listing advertisements adapted to displayed contents, and another example is targeting advertisements based on user behavior displayed based on web history.

These are core approaches to web advertisement, which comprise strategy and analysis to effectively pass product information to target users. As a benchmark indicating the importance of web marketing, the amount of advertising expense in Japan has grown 657 times from 1995 to 2014 [1]. However, the growth rate has remained unchanged recently. This is because the component technology is not making any significant changes in the amount of investment in advertising as well as business type of advertising.

Several kinds of web services/Social Network Service (SNS) utilize registered individual information and service usage history to display personalized advertisements [2] . However, these services require active registration of personal information, and it is difficult to personalize advertising without entering personal information. Behavioral targeting advertising utilizes cookies generated on each website for personalization purposes; however, the stored information may be outdated and might therefore not reflect user's current interests.

User interests can be roughly classified into the following two types: continuous long-term interests regardless of the time or period such as hobbies and intentions, and short-term interests which last for a certain period of time and are focused on current tasks or investigations. Therefore, we focus on the information that the user has acquired most recently in order to predict the user interest and the transition of the user interest. By using our proposed method, we expect applications such as conventional web marketing, advancement of web advertisement, and dynamic design of information system to be based on prediction of user behavior.

It is assumed that the short-term interest of the user is limited to the latest 30 minutes [3], and the acquired web content by each user will include current interest information within this period. For feature analysis of each user interest, supervised learning is used for predicting user interest by regression of user acquired information. In this paper, the data used for learning is online behavioral history collected through the clicked log of web advertisements on several real web services. The effectiveness of the proposed method is evaluated in comparison with the general method that uses the statistical method of words embedded in web content.

Section 2 describes and discusses how to acquire information indicating user interest to the analysis based on web activities. Section 3 clarifies the proposed way of analyzing this information. Section 4 evaluates the proposal, and section 5 gives consideration. Section 6 concludes this paper.

II. USER INTERESTS

This section discusses how information related to current user interests can be determined from their web behavior.

A. Related Research

Regarding related research, Siriaraya et al. [4] proposes a method for analyzing the potential interest of the long-term/short-term user based on the analysis of user activities on the web in the same way as it is done in this paper. Feature analysis compares the Fully Qualified Domain Name (FQDN) of the site visited by the user and the category of the website. It is clear that the method using website category

 ISBN: 978-1-61208-736-8

has better performance, and the long and short analysis period has no relation to prediction performance. Compared with the proposed method, which uses web content for analysis, the analysis load of learning features based on FQDN and website category involves less processing load caused by no text processing than the proposed method. However, using this information for feature analysis only makes it possible to classify website features. In other words, it only predicts the classification of user interest information. On the other hand, the method proposed in this paper analyzes the information contained in the content as a feature although the processing load increases. It is possible to make judgment based on similarity considering the content in more detail than the classification of interest information. It is also clarified that the prediction performance can be expected to improve by introducing Recurrent Neural Network (RNN) to consider the order of browsing history. The use of order reveals the context of browsing history. And the effect of emphasizing the interest information of short–term users can be expected compared to the conventional method.

Van den Poel et al. [5] proposes a method that predicts consumer behavior from the behavior before and after making a purchase through an EC site. In particular, a model based on the logit model is proposed. However, since the method is based on statistical prediction, it cannot predict user interest. It is impossible to predict behavioral patterns that have not occurred before. Since this paper analyzes the information of the acquired content itself, it is possible to predict user interest information from the viewpoint of similarity and correlation among contents.

B. Information Indicating User Interests

This research assumed that user interests at each point in time are included in information retrieved around that time. By analyzing the acquired information and its characteristics, current user interests could be elucidated.

As such interests change over time, it was assumed that the same interest lasted up to 30 minutes at most. Therefore, user web behavior history characteristics ware used to derive user interests. Fig. 1 shows the web behavioral history of users at each point in time ($t = 0$).

C. Web Behavioral History

The web behavioral history URLs used in this study were collected by executing a script when acquiring web content, after which information was stored with the user ID generated from the first access and time on the database. The data were then analyzed for user interests by automatically scraping the content from the collected URLs for each user using PhantomJS [9] and Selenium [10] headless browser, which sequentially accessed each URL to acquire content.

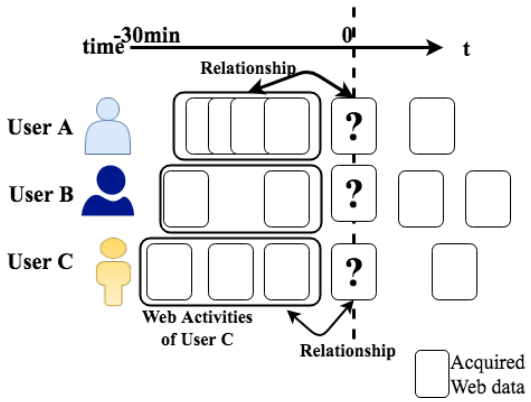

Figure 1: User Interests

D. Extracting the Main Content

There are numerous methods for extracting main content, e.g., presumption based on learning the web content context using the natural language method [6] or learning content placement within a webpage using machine learning [7].

Yamamoto et al. [8] assumed that the main content had the most sentences (number of characters). The main content occupied an average of 78.8% of all texts in the web content, which was the largest ratio obtained from the analysis of real web pages. Therefore, in this paper, it was assumed that the main content locations were based on the ratio of block size to the entire text volume from <body> tag to </body> tag. Therefore, it was only necessary to count the characters on the web page, which lessened the extraction load compared to learning the web context.

E. Deriving Information from the Content

Almost all targeted web contents were written in Japanese. Therefore, it was necessary to analyze each word in order to extract the characteristics; note that some words were combined with their postpositional particles or auxiliary verbs. Therefore, unlike the extraction of English words, it was necessary to analyze the words in texts using morphological analysis, after which the words unrelated to the information context such as particles, conjunctions, or numerals were removed to avoid performance degradation and reduce processing load. This study used the open source segmentation library MeCab [11] for morphological analysis.

III. Model Construction

This section presents a learning model, which is constructed on the basis of the information extracted from each user. Fig. 2 provides an overview of the proposed method.

A. Data for Learning

The web information acquired by the user, who clicks the web advertisement as shown in Fig. 2, is used as the data for learning. As shown in Section 1, the current web advertisement is selected as the advertisement which has an attribute related to or similar to that of the web content

 ISBN: 978-1-61208-736-8

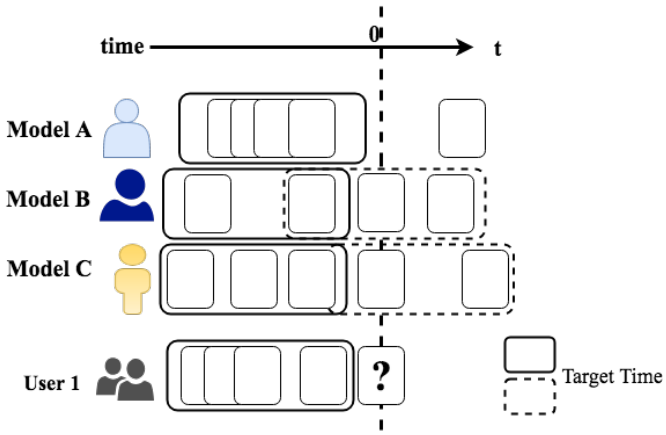

Figure 2: Overview of Constructing Models

for the ad space in web pages. Delivered advertisements are dynamically determined by bidding for advertisements that are adapted to the attributes of acquiring user (static information such as age and living area, dynamic information such as past web history). Therefore, the user who clicks on the advertisement displayed on the web content could be considered as currently interested in the web advertisement. By increasing the number of target data, it can be considered that the number of users who are more interested in web advertisements and clicked on them increases than the users who accidentally clicked on the web advertisements. Therefore, the acquired information from the users who clicked this web advertisement can be considered as the indication of their current interests, and is used as the data for learning.

B. Context Learning

A paragraph vector–distributed memory (PV–DM) model [12] was used to learn the order of the words included in the information. At the time, the PV–DM model determines the learning contexts in the information as the words in the window are treated as input, and the next word is adjacent to the neural network window, as shown on the left side of Fig. 3.

In Fig. 3, the parameter, “window size,” is set to three words. In an environment where the parameter “window size = 3 words” is set, the information context learning proceeds by designating “wordA” ∼” word C” in the window as the information contexts in order to output “target word D.” The output is then learned by moving the window for all words included in the information. By sliding the target to the right, the learning proceeds by designating “word B” ∼ “word D” as the contexts in order to predict “word E.” The proposed method also learns in the opposite direction; that is, by designating the input window on the right side in Fig. 3 (“wordE” ∼ “word G”), the output is the previous word (“target word D”). In short, the proposed method

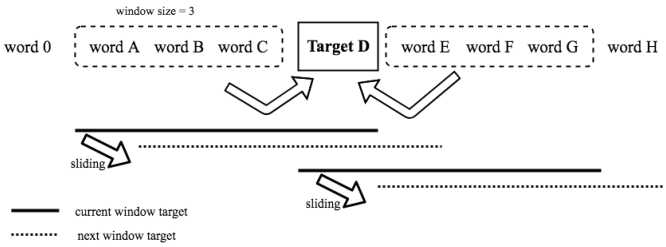

Figure 3: Context Learning

learns on a neural network so that it can predict words on both sides of the defined window parameter “window size,” which enables effective learning even if only a few words are included the window. Based on these processes, the learning model is structured by using the derived information, from which it forms the learned word vectors and information vectors.

C. Prediction

In order to predict the user interest, vectorization of the latest acquired information is executed by using the constructed model as in the case of the data for learning. The vectorized acquisition information of each user includes the feature of their interest. (corresponding to the “?” of User 1 in Fig. 1).

Therefore, the predicted information is regressed to a vector indicating a similar feature. However, such vector information is not useful as it is for the predicted information. Therefore, the candidates of predicted information (e.g., the web advertisement and web content) are vectorized in advance. The similarity between the vectorized candidates of predicted information and the vectorized users interest is derived, and the one closest candidate information to the user interest is taken as the predicted information. In other words, the most similar information is determined by the cosine similarity, which measures the angle formed by two vectors.

IV. Evaluation

This section evaluates the proposed method for determining the user interests from user behaviors acquired from the real web services.

A. User Behaviors on the Web

The user behavior extracted for the evaluation is clicked as the log of web advertisement mentioned in Section 3.1 as for learning and evaluating, which is collected from various unspecified web services on various domains. The data used in this paper use the log of web advertisement notified to the advertisement network. It is information shared among the advertisement frames (media) for displaying the recommendations and advertisements within the web content, called the advertisement network.

 ISBN: 978-1-61208-736-8

It is not located at a specific domain, but at unspecified websites on various domains. The data included user IDs and the URL entries for the executed sites. The user ID was generated randomly when each user first accessed any of the corresponding websites. Then, the same user ID was used when the same script was executed in any of their websites; that is, the user web behaviors were traced using this ID. This evaluation used about $5,000$ user behaviors (the number of user was $1,300$) for the learning, and another 150 user behaviors for the evaluation test.

B. Evaluation Scenario

In this evaluation, the model was constructed based on the user behaviors, as shown in TABLE I, after which the evaluation data were vectorized to predict the user interests and derive the most similar behaviors. Then, the proposed method and a method using the statistical properties of words and web content were compared. The expression of statistical properties for appearing words is done by the LDA method [14], in which the use for probabilistic reduction of vector dimensions in the tf–idf [13] method weights the appearing words based on frequency.

To demonstrate the effectiveness of main content extraction, the performance comparison was also evaluated. TABLE II shows the development environment.

TABLE I. USER BEHAVIORS ON THE WEB

Environment	Specification
Number of User Behaviors	4500
Number of Users	1300
Number of Users for test	150User

TABLE II. DEVELOPMENT ENVIRONMENT

Environment	Specification
Pre-Processing	gensim [15]
Modeling	Tensor Flow 1.8 [16], Python 3.6.2
OS	Ubuntu 16.10
System	Docker 17.09

C. Results

Fig. 4 shows the evaluation results, in which the $x-$axis is the precision rate, and the $y-$axis is the recall rate. In the graph, "Proposal (only main)" refers to the method when only the main content was used for the modeling, "Received all data" refer to the method when all acquired web content was used for the modeling, and "Conventional" was the statistical method based on the word frequency including the main content. In general, the graph on the upper right side of the graph shows a superior performance for the recall–precision rate, indicating that the proposed method was able to effectively express the information characteristics.

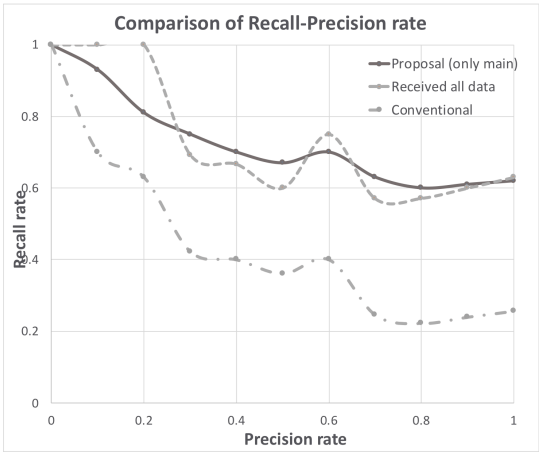

Figure 4: Comparison of R–P rate

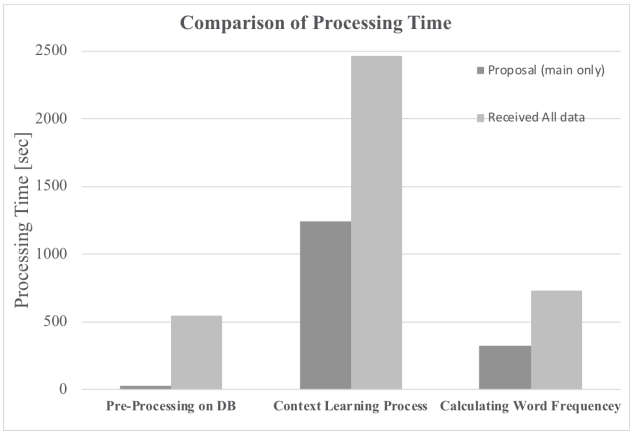

Figure 5: Processing Time

The "Received all data" case performed better than the "Proposal (only main)" case in some intervals, which was possibly due to the partial information loss. The main content characteristics were emphasized in the embedded web advertisements, which were similar to the main content in the sides and the header. The "Received all data" case dynamically generated the web advertisements and "search engine optimization" tags as well as the main content, which was based on the main content similarities and web history. Therefore, it was concluded that the "Received all data" performance was unstable, but the "Proposal (only main)" was stable at some intervals.

Fig. 5 shows the processing time comparisons for extracting and not extracting the main content from the web content. Specifically, it shows the time taken for each process from scraping to the extraction of the words from the contents on the DB; that is, as this difference was the difference in the information volume to be processed, accordingly, the impact increases in proportion to the target information volume.

 ISBN: 978-1-61208-736-8

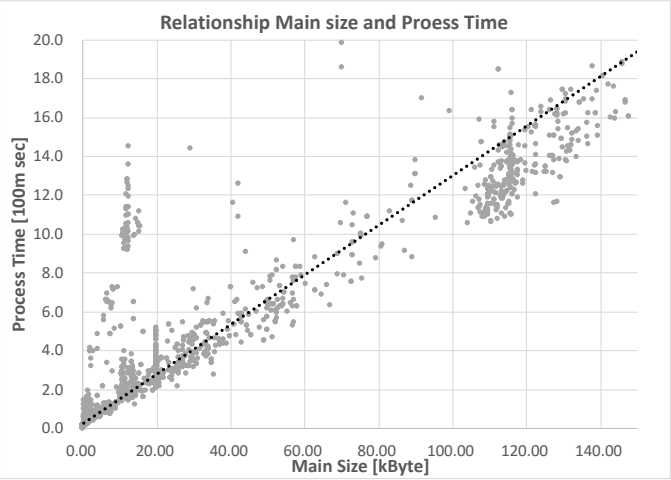

Figure 6: Processing time and Data Volume

V. CONSIDERATION

Based on the results, this section examines the effectiveness of the proposed method.

A. Based on the evaluation

Based on the evaluation results, the proposed method was found to be able to identify similar information related to the users web interests. In particular, extracting the main content by using the learned model rather than the statistical information showed a better and more stable performance. Due to the assumption that the acquired information was biased toward a specific field or topic (lack of comprehensive information), it was concluded that the conventional method had a poorer performance because of insufficient data as the $4,500$ items in TABLE I were inadequate for the method characteristics.

The proposed method only targeted the web content acquired by the users; therefore, the performance was more stable and the processing load was smaller as the main contents were previously extracted (this process corresponded to data cleansing). However, to predict the user interests, the following processes were necessary: acquiring the most recent information, extracting the main contents, and analyzing these contents; whereas, the conventional method use cookies. If the proposed method were to be applied to a real system, the key problems would be at the respective nodes, where these processes were to be executed.

The load processing for extracting the main content from each web page was then clarified for the $3,000$ randomly extracted websites by calculating the time for the pre–processing and word extraction from the main contents. Fig. 5 shows the relationship between processing times and targeted data volume.

In Fig. 6, the $x-$axis shows the size of the main contents [kByte], and the $y-$axis shows the extraction time for the main content and words, for which the average size of the main content to page size was 46.53%; that is the main content occupied 46% of the user's acquired web contents. In addition, the main content was less than 150[kByte] for $2,800$ of the websites, which was a majority of the $3,000$ websites. Fig. 5 only plots this range on the graph, from which it can be seen that the processing time linearly increased with a slope of 0.13 in proportion to the main content size. However, to implement this method, the load balancing would need to be primarily dealt with as the load increases depending on the processing and extracting processes.

If it is assumed that each node simultaneously executed these processing steps with content acquisition, then it will take about 1 [sec] to process the average data size for each content; therefore, the processing load could be distributed to the nodes, leaving a possibility that the web usability of each user could be lost.

Therefore, for effective implementation, it would be necessary to design a system model which is focused on how this processing could be executed.

B. Future Work

This section discusses the possible improvements to the proposed method. The model based on the consideration of the user web behavioral history was able to effectively extract the information related to the user interests; therefore, the users who clicked or watched a displayed web advertisement corresponding to their interests could be targeted. As these users showed an interest in the display advertisement, these could be seen to have relevance to the user latest web activities and could be used as a part of the users' information history.

The user interest duration was assumed to be up to 30 minutes in this paper. However, this could change depending on the situation and the people. Accordingly, it could be possible to determine the user interest displacement and thought transitions from a correlation of the obtained web data. Therefore, a method could be designed to detect the user interest transition from the acquired data correlations, which could remove the non–target data from the training data and improve the quality and effectiveness of the proposed method.

The data used in this evaluation were focused on the user web behavioral history acquired from a web service on one day; however, it is necessary to evaluate this by extending the duration. If the proposed method were to be implemented on a real system, it would be necessary to solve the load problem for the acquisitions and analysis of the user web behavioral history data; therefore, a method to distribute this through networks could also be considered.

VI. CONCLUSION

In this paper, we assume that a web user latest interests would be included in their web behavioral history within the

 ISBN: 978-1-61208-736-8

latest 30 minutes of web browsing. Therefore, a prediction method was proposed based on the analysis and extraction of the users' characteristics. The proposed method was found to more easily identify the latest user interests than the conventional method based on cookies. The evaluation also indicated that the proposed method was better able to predict the user interests than the current method based on the statistical information. Future works will improve the model performance to better categorize the user interests and develop a method to dynamically detect the interest duration.

ACKNOWLEDGMENT

This work was partly supported by MEXT KAKENHI Grant Number 17K01149

REFERENCES

[1] Dentsu, "Advertising Expending Reports in Japan", http://www.dentsu.com/knowledgeanddata/ad_expenditures/ , Retrieved September 8, 2019

[2] S. Hirose, "Textbook for Advertisement technology", Shoeisya, 2016 (written in Japanese)

[3] Y. Watanabe and Y. Ikegaya, "Effect of intermittent learning on task performance: a pilot study", Journal of Neuronet, Vol.38 pp.1–5, 2017

[4] P. Siriaraya, Y. Yamaguchi, M. Morishita, Y. Inagaki, R. Nakamoto, J. Zhang, J. Aoi, and S. Nakajima, "Using categorized web browsing history to estimate the user's latent interests for web advertisement recommendation", 4429-4434. 10.1109/BigData.2017.8258480

[5] V. den Poel and B. Wouter, "Predicting Online-purchasing Behaviour", European Journal of Operational Research 166, pp.557–575, 10.1016/j.ejor.2004.04.022

[6] B. Orkut, G. Hector, and P. Andreas, "Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices" , Proc. 10th International Conference on World Wide Web, pp.652–662, 2001

[7] K. Nakamura, S. Tanaka, Y. Yamamoto, and S. Abiko, "Method of Filtering Harmful Information Considering Extraction Range of Word Co-occurrence", IPSJ Journal, Vol.54, No.2, pp.571–584, IPSJ, 2013

[8] Y. Yamamoto, K. Nakamura, S. Tanaka, and S. Abiko, "Proposal Research of Web Page Segmentation Method for Extracting and Describing Each Article in Detail", IPSJ Journal, Vol.55 No.2, pp.874–891, 2014

[9] PhantomJS,"http://phantomjs.org/", Retrieved September 8, 2019

[10] Selenium,"https://www.seleniumhq.org/", Retrieved September 8, 2019

[11] MeCab, "http://taku910.github.io/mecab/", Retrieved September 8, 2019

[12] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents" , Proc. of Int. Conf. on Machine Learning, pp.1188–1196, 2014

[13] C.D. Manning, P. Raghavan, and H. Schutze, "Scoring, term weighting, and the vector space model", Introduction to Information Retrieval. pp100–123, 2008

[14] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research, pp. 1107–1135, 2003

[15] topic modelling for humans, "https://radimrehurek.com/gensim/", Retrieved September 8, 2019

[16] TensorFlow, "https://www.tensorflow.org/", Retrieved September 8, 2019

 ISBN: 978-1-61208-736-8