Mining Life Events Based on the Fluctuation of Users' Web Access

Hao NIU[†] Mori KUROKAWA[†] Shigeru KUROYANAGI[‡] and Arei KOBAYASHI[†]

† KDDI Research, Inc. 3-10-10 Iidabashi, Chiyoda-ku, Tokyo, 102-8460, Japan
‡ Supership Inc. 5-4-35 Minami-Aoyama, Minato-ku, Tokyo, 107-0062, Japan
E-mail: † {ha-niu, mo-kurokawa, kobayashi}@kddi-research.jp, ‡ shigeru.kuroyanagi@supership.jp

Abstract The life events are regarded to play an important role in consumers' purchasing behaviors. In case such information is available, advertisement recommendation systems can accomplish more effective advertising. However, this kind of privacy information is difficult to be obtained directly and data mining-based prediction is usually necessary. In this paper, a study on predicting users' life events in the near future based on the fluctuation of their web access is conducted. The life event related websites are first specified by exploiting users' web access logs. Then, the prediction model is developed based on users' fluctuation of accessing these websites. The proposed scheme is evaluated on a data set which is comprised of the web access logs of thousands of anonymous users within two years, and the life event-house moving is taken as an example. The results show that the higher precision and recall can be achieved by our developed model compared to the random guess.

Keyword Web Access, Life Event, E-commerce

1. Introduction

People nowadays are increasingly relying on the Internet for information searching and shopping. The Internet users not only actively use the search engines and E-commerce websites, but also receive the information passively from the advertisement recommender systems. If the distributed advertisements from these systems provide the information expected by the users, the clickthrough rate can be significantly improved. To distribute such correct advertisements, the advertisement recommender systems should distinguish who are the target users. There are mainly two approaches to find target users: content-based filtering [1] and collaborative filtering [2]. Both of the two approaches attempt to mine users' profiles or similarities from different kinds of data.

In users' profiles, the life events (e.g., house moving, marriage) affect consumers' purchasing behaviors very much. Users facing life events prefer to search for related information and purchase related products or services in advance. For example, house-moving users tend to find the information about rental houses and order some pieces of furniture in advance, while the newlyweds tend to find the information about wedding ceremony and buy wedding gowns ahead. If the occurrence of users' life events in the near future is known, it is possible for the advertisement recommender systems to distribute related advertisements precisely.

However, the life events of users are usually privacy information and cannot be obtained directly. Mining life events based on users' data for E-commerce has been studied in recent research works. In [3], users' shopping history in the China's largest E-commerce website, Taobao, is utilized to analyze the Mum-Baby status of users with the Maximum Entropy Semi Markov Model. In [4], both the prediction and product recommendation approaches regarding the Mum-Baby status are studied using the open data of shopping history from Tianchi, a data science and machine learning competition website of Alibaba. Different from the above two works using shopping history, the app adoption of mobile devices and twitter logs are adopted separately in [5] and [6] to predict the current or near-future life stages (events) of users.

There have been also some trials predicting users' life events based on the web access in industry [7-9]. In [7-8], keywords are first extracted from users' access data, and then these keywords are compared with the predefined life event related keywords to predict the occurrence of life events. The prediction approach in [9] is similar, except that the websites(URLs) instead of keywords are utilized. Both works in [7-8] and [9] require a predefined database of life event related keywords or websites, but how to generate this kind of database is not explained. Also, these approaches are simple rule-based, which may need rich experience of operators and be difficulty to maintain the rules. To solve these problems, a proposal to specify the life event related websites is first designed by us, and then a machine learning prediction model for the occurrence of life events based on users' fluctuation of accessing these related websites is developed. A data set, which is comprised of the web access logs of thousands of anonymous users within two years, is utilized to evaluate our scheme. The life event-house moving is taken as an example, and the prediction results show that the higher precision and recall can be achieved by our developed model compared to a random guess method.

2. Life event prediction proposals

Each life event is assumed to have its related websites. For example, house moving is related to rental house and furniture websites, marriage is related to wedding ceremony and wedding gown websites, and so on. The access of these related websites is usually increased when a user will experience or is experiencing a life event. Based on the above assumption, we propose a scheme for life event prediction, the framework of which is shown in Fig 1.

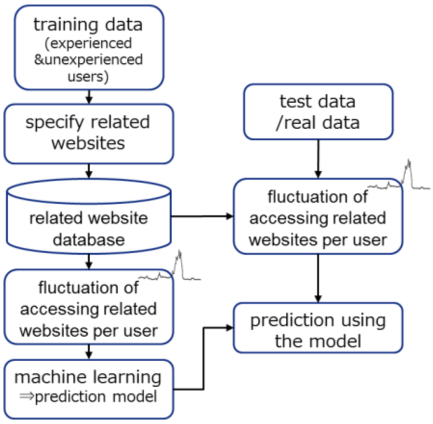
training data
(experienced &unexperienced users)
specify related websites
related website database
fluctuation of accessing related websites per user
machine learning ⇒prediction model
test data /real data
fluctuation of accessing related websites per user
prediction using the model

Fig 1. Framework of the scheme for life event prediction

Let's denote the users who experienced the life event as experienced users and the users who did not experience the life event as unexperienced users in the training data. First, the statistical difference of the experienced users before and after the life event, as well as the statistical difference between the experienced users and the unexperienced users are exploited to specify the life event related websites, which are then saved in a database. Next, the fluctuation data of accessing these related websites before the life event is calculated per user in the training data as the features. The prediction model is then developed by using machine learning on these fluctuation data and the life event labels. For the evaluation on test data or the prediction in real environments, users' fluctuation data of accessing related websites are similarly calculated according to the related website database, which is inputted into the developed model to predict whether users will experience a life event or not in the near future. The details are described as follows.

2.1 Specifying life event related websites

For machine learning, the statistical difference between positive samples and negative samples (the experienced users and the unexperienced users in this paper) is usually utilized to find the significant features. However, the life events are generally temporary, such that the experienced users themselves can be treated as the unexperienced users when they are not during the period of life events. The statistical difference of the experienced users during and not during the life events is also able to be utilized to extract the significant features. Both of the above two types of statistical difference are considered by us to specify the life event related websites.

Type A: Statistical difference of the experienced users during and not during the life events for specifying the life event related websites

The information on the period of accessing life event related websites is not available in general, and thus it is not intuitive to decide when is the start point to find the statistical difference. Therefore, we adopted a backward method, which compares the web access logs before and after the life event, to specify the related websites. The proportion of users who accessed a website in the experienced user group during a time window T is aggregated as Table 1. The length of T can be adjusted, e.g., one week or one month. T_0 indicates the time window in which the life event occurs, while T_n indicates the time windows before (after) T_0 when n is a negative (positive) number. The range of n can be decided as needed, and [-5,1] is adopted in Table 1. $P_{m,n}$ means the proportion of users who accessed a website in the experienced user group, for the website$m$ and time window T_n. For example, when T_0 indicates 2017.12, T_{-1} and T_1 indicates 2017.11 and 2018.01 respectively. If 100 experienced users access the web1 in 2017.12 and there are 500 experienced users in total, $P_{1,0}$ is 100/500 = 0.2.

Table 1. The proportion of users who accessed a website in the experienced user group

	T_{-5}	T_{-4}	T_{-3}	T_{-2}	T_{-1}	T_0	T_1
web1	$P_{1,-5}$	$P_{1,-4}$	$P_{1,-3}$	$P_{1,-2}$	$P_{1,-1}$	$P_{1,0}$	$P_{1,1}$
web2	$P_{2,-5}$	$P_{2,-4}$	$P_{2,-3}$	$P_{2,-2}$	$P_{2,-1}$	$P_{2,0}$	$P_{2,1}$
web3	$P_{3,-5}$	$P_{3,-4}$	$P_{3,-3}$	$P_{3,-2}$	$P_{3,-1}$	$P_{3,0}$	$P_{3,1}$
…							

Since we use the backward technique, the ratio of $P_{m,n}$ (n <0) to $P_{m,1}$ is calculated:

$$RA_{m,n} = P_{m,n}/P_{m,1}$$

Two threshold values γ_1 and γ_2 are determined to specify the related websites. First the column of T_{-1} is checked and the websites with ratio $RA_{m,-1} \geq \gamma_1$ are selected as the related websites. The set of these websites and its cardinality (the number of these websites) are denoted by $\boldsymbol{A}_{-1}$ and N_{-1} respectively. Then, the column of T_{-2} is checked similarly and the websites with ratio $RA_{m,-2} \geq \gamma_1$ are selected as the related websites. The set of these websites and its cardinality are denoted by $\boldsymbol{A}_{-2}$ and N_{-2}. The similar process is conducted for T_{-3} and so on. At last, the process stops at a value s with $N_s < \gamma_2$, or n reaches to a predefined minimum value s_{min} (in this case s is set to be s_{min}). s+1 is regarded as the start point of accessing the life event related websites. The union of sets, $\boldsymbol{A} = \boldsymbol{A}_{s+1} \cup \boldsymbol{A}_{s+2} \cup \ldots \cup \boldsymbol{A}_{-2} \cup \boldsymbol{A}_{-1}$, is the acquired set of related websites for *Type A*.

Type B: Statistical difference between the experienced users and the unexperienced users for specifying the life event related websites

The statistical difference between the experienced users and the unexperienced users is only considered from T_{s+1} to T_{-1} (s was derived above). The proportion of users who accessed a website in the unexperienced user group during a time window is aggregated as Table 2.

Table 2. The proportion of users who accessed a website in the unexperienced user group

	T_{s+1}	T_{s+2}	…	T_{-2}	T_{-1}
web1	$P'_{1,s+1}$	$P'_{1,s+2}$	…	$P'_{1,-2}$	$P'_{1,-1}$
web2	$P'_{2,s+1}$	$P'_{2,s+2}$	…	$P'_{2,-2}$	$P'_{2,-1}$
web3	$P'_{3,s+1}$	$P'_{3,s+2}$	…	$P'_{3,-2}$	$P'_{3,-1}$
….					

The ratios of $P_{m,n}$ in Table 1 to $P'_{m,n}$ in Table 2 are then calculated:

$$RB_{m,n} = P_{m,n}/P'_{m,n}$$

For each column from T_{s+1} to T_{-1}, the websites with ratio $RB_{m,n} > \gamma_1$ are selected as the related websites, the set of which are denoted by $\boldsymbol{B}_n$. The union of sets, $\boldsymbol{B} = \boldsymbol{B}_{s+1} \cup \boldsymbol{B}_{s+2} \cup \ldots \cup \boldsymbol{B}_{-2} \cup \boldsymbol{B}_{-1}$, is the acquired set of related websites for *Type B*.

$\boldsymbol{A}$ and $\boldsymbol{B}$ are thought to be almost the same when using ideal data set. However, the real data set generally has limited user number and limited records, which makes $\boldsymbol{A}$ is different from $\boldsymbol{B}$ in the real environments. Different combinations of $\boldsymbol{A}$ and $\boldsymbol{B}$ (only $\boldsymbol{A}$, only $\boldsymbol{B}$, $\boldsymbol{A} \cup \boldsymbol{B}$ and $\boldsymbol{A} \cap \boldsymbol{B}$) can be considered as the final set of related websites, the differences of which will be illustrated in the experiment section.

2.2 Developing the prediction model based on the fluctuation of users' web access

The fluctuation of accessing the related websites per user (the experienced users and unexperienced users) is used as the features. Specifically, the proportion of records accessing the related websites to all the records per user is aggregated first for each time window from T_s to T_{-1}. Let's denote $N_{i,n,r}$ and $N_{i,n,t}$ are the number of recodes accessing related websites and the number of all the records respectively for user$i$ in the time windows T_n. The proportion is derived as $p_{i,n} = N_{i,n,r}/N_{i,n,t}$. Then the difference (fluctuation) $F_{i,n}$ of the proportions between T_n and T_{n-1} is calculated per user for $n \in \{s+1, s+2, \ldots, -2, -1\}$, i.e, $F_{i,n} = p_{i,n} - p_{i,n-1}$.

If user$i$ is an experienced user who moved in the time window T_0, the label is set to be 1. Otherwise, the label is set to be 0. Then the Table 3 can be obtained, which is used to perform the machine learning to develop a prediction model. For test data or real data, the occurrence of life events in the near future (i.e., in next month here) can be predicted after processing users' web access logs in the same form.

Table 3. Features (fluctuations) and labels for learning the life event prediction model

	Features					label
	T_{s+1}	T_{s+2}	…	T_{-2}	T_{-1}	
user1	$F_{1,s+1}$	$F_{1,s+2}$	…	$F_{1,-2}$	$F_{1,-1}$	1
user2	$F_{2,s+1}$	$F_{2,s+2}$	…	$F_{2,-2}$	$F_{2,-1}$	1
user3	$F_{3,s+1}$	$F_{3,s+2}$	…	$F_{3,-2}$	$F_{3,-1}$	0
….						

3. Experiments

The experiment evaluation is performed on a data set, which is comprised of the web access logs of thousands of anonymous users within two years. The life event-house moving is considered, and the house moving labels are obtained from questionnaires. Also, the URL's network location part (netloc) is adopted to identify the websites.

3.1 Data preparation

All the users are divided into two groups according to whether they moved or not: moved users and unmoved users. The parameters T, γ_1, γ_2 and s_{min} are set to be one month, 2, 1 and -6 initially. Since the data of T_1 is necessary for obtaining the related website set $\boldsymbol{A}$, the users who moved in the 24th month are not considered. Also, the data from $T_{s_{min}}$ to T_{-1} is necessary to obtain the features, the users who moved from 1st to 6th months are not considered either. Therefore, only the moved users who moved from the 7th to 23rd months are considered.

The data is prepared as Table 4, in which the samples of the N^{th} month (N is from 7 to 23) are prepared based on the (N-6)$^{\text{th}}$ ~ (N+1)$^{\text{th}}$ months' logs of the users who moved in the N^{th} month and all the unmoved users. However, the users who have no web access logs from the (N-5)$^{\text{th}}$ to (N-1)$^{\text{th}}$ months are excluded.

Table 4. The data preparation[1]

		1	2	3	4	5	6	7	8	9	10	11	…
7	users moved in 7th month	🟦	🟦	🟦	🟦	🟦	🟦	🟧	🟦				
	unmoved users	🟦	🟦	🟦	🟦	🟦	🟦	🟧	🟦				
8	users moved in 8th month		🟦	🟦	🟦	🟦	🟦	🟦	🟧	🟦			
	unmoved users		🟦	🟦	🟦	🟦	🟦	🟦	🟧	🟦			
9	users moved in 9th month			🟦	🟦	🟦	🟦	🟦	🟦	🟧	🟦		
	unmoved users			🟦	🟦	🟦	🟦	🟦	🟦	🟧	🟦		
…													

After having prepared the data, the total samples are aggregated together and arranged randomly. Then the samples are split into 10 folds for cross validation.

3.2 Specifying house moving related websites

For every iteration, the house moving related websites are specified using the training data according to the last section. Corresponding to Table 1, a fragment of the proportion of users who accessed a website in the moved user group is shown in Table 5(a), from which the ratio of $P_{m,n}$ (n <0) to $P_{m,1}$ is calculated in Table 5(b).

Table 5. (a) The proportion of users who accessed a website in the moved user group $P_{m,n}$

	T_{-5}	T_{-4}	T_{-3}	T_{-2}	T_{-1}	T_0	T_1
web1	0	0	0	0	0.004	0	0
web2	0	0	0	0	0	0	0.004
web3	0	0	0	0.004	0.008	0.008	0.004
web4	0	0.005	0	0	0	0	0
web5	0	0	0	0	0.004	0	0.004
…							

(b) The ratio of $P_{m,n}$ (n <0) to $P_{m,1}$

	T_{-5}	T_{-4}	T_{-3}	T_{-2}	T_{-1}
web1	0	0	0	0	∞
web2	0	0	0	0	0
web3	0	0	0	1	2
web4	0	∞	0	0	0
web5	0	0	0	0	1
…					

The ratios in Table 5(b) are checked from T_{-1} to T_{-5} to obtain the sets of the relate websites for *Type A*, i.e., $\boldsymbol{A}_{-1},\boldsymbol{A}_{-2},\boldsymbol{A}_{-3},\boldsymbol{A}_{-4},\boldsymbol{A}_{-5}$. For example, in the time window T_{-1}, $P_{1,-1}/P_{1,1} = \infty \geq \gamma_1 = 2$ and $P_{3,-1}/P_{3,1} = 2 \geq \gamma_1 = 2$, website1 and website3 are inserted into $\boldsymbol{A}_{-1}$. In our data set, the cardinalities of the obtained $\boldsymbol{A}_{-1,}\boldsymbol{A}_{-2},\boldsymbol{A}_{-3},\boldsymbol{A}_{-4},\boldsymbol{A}_{-5}$ are all larger than or equal to $\gamma_2 = 1$. Thus, s equals to the predefined minimum value -6, and s+1=-5 is regarded as the start point of accessing the house moving related websites. $\boldsymbol{A} = \boldsymbol{A}_{-5} \cup \boldsymbol{A}_{-4} \cup \boldsymbol{A}_{-3} \cup \boldsymbol{A}_{-2} \cup \boldsymbol{A}_{-1}$ is utilized to obtain the total related websites for *Type A*.

Next, the proportion of users who accessed a website in the unmoved user group from T_{s+1} (T_{-5}) to T_{-1} is calculated as Table 6(a), and thus the ratio of $P_{m,n}$ to $P'_{m,n}$ can be obtained as Table 6(b).

Table 6. (a) The proportion of users who accessed a website in the unmoved user group $P'_{m,n}$

	T_{-5}	T_{-4}	T_{-3}	T_{-2}	T_{-1}
web1	0.0002	0.0002	0.0011	0.0021	0.0029
web2	0.0004	0.0008	0.0012	0.0013	0.0014
web3	0.0014	0.0015	0.0021	0.0024	0.0032
web4	0.0000	0.0000	0.0000	0.0000	0.0000
web5	0.0001	0.0002	0.0003	0.0007	0.0009
….					

(b) The ratio of $P_{m,n}$ to $P'_{m,n}$

	T_{-5}	T_{-4}	T_{-3}	T_{-2}	T_{-1}
web1	0	0	0	0	1.38
web2	0	0	0	0	0
web3	0	0	0	1.67	2.5
web4	0	∞	0	0	0
web5	0	0	0	0	4.44
…					

The ratios are also checked from T_{-1} to T_{-5} (or from T_{-5} to T_{-1}). For example, in the time window T_{-1}, $P_{3,-1}/P'_{3,-1} = 0.008/0.0032 = 2.5 \geq \gamma_1 = 2$, $P_{5,-1}/P'_{5,-1} = 0.004/0.0009 = 4.44 \geq \gamma_1 = 2$, thus website3 and website5 are inserted into $\boldsymbol{B}_{-1}$, and so on. Finally, $\boldsymbol{B} = \boldsymbol{B}_{s+1} \cup \boldsymbol{B}_{s+2} \cup \ldots \cup \boldsymbol{B}_{-2} \cup \boldsymbol{B}_{-1}$ is calculated to obtain the total related websites for *Type B*. The final sets of the related websites by different combinations of $\boldsymbol{A}$ and $\boldsymbol{B}$ (only $\boldsymbol{A}$, only $\boldsymbol{B}$, $\boldsymbol{A} \cup \boldsymbol{B}$ and $\boldsymbol{A} \cap \boldsymbol{B}$) are adopted in our experiments later.

3.3 Prediction model development and evaluation

The proportion of records accessing related websites to all the records per user, $p_{i,n}$, is calculated for both the training data and the test data, which is shown in Table 7(a) ($\boldsymbol{A} \cup \boldsymbol{B}$ is adopted here as an example). Based on the table of $p_{i,n}$, the features for developing the prediction model are obtained by $F_{i,n} = p_{i,n} - p_{i,n-1}$ in Table 7(b).

After attaching the labels, the prediction model is developed by using logistic regression classifier. The

[1] The data of the (N+1)$^{\text{th}}$ month is only used for obtaining the related website set $\boldsymbol{A}$.

threshold for predicting whether an example is positive or negative is adjusted to maximum the mean precision, when further splitting the training data into 10 parts and treating each part as the validation set. The random guess method is adopted as the baseline like [5]. Fig 2 illustrates the performance (precision and recall) improvement of our developed model for $\gamma_1 = 2$ and $\gamma_1 = 10$ respectively ($\gamma_2 = 1$), where different combinations of $\boldsymbol{A}$ and $\boldsymbol{B}$ (only $\boldsymbol{A}$, only $\boldsymbol{B}$, $\boldsymbol{A} \cup \boldsymbol{B}$ and $\boldsymbol{A} \cap \boldsymbol{B}$) are considered.

Table 7. (a) The proportion of records accessing related websites to all the records per user

	T_{-6}	T_{-5}	T_{-4}	T_{-3}	T_{-2}	T_{-1}
Training data						
user1	0	0	0	0.279	0.362	0.428
user2	0.514	0.509	0.71	0.325	0.442	0.469
user3	0	0	0	0	0	0
…						
Test data						
user1	0.154	0.893	0.111	0.143	0.68	0.8
user2	0	0.313	0.267	0.26	0.001	0.044
…						

(b) Features (fluctuations) and labels for developing the prediction model

	Features					label
	T_{-5}	T_{-4}	T_{-3}	T_{-2}	T_{-1}	
Training data						
user1	0	0	0.279	0.083	0.066	0
user2	-0.005	0.201	-0.385	0.117	0.027	1
user3	0	0	0	0	0	0
…						
Test data						
user1	0.739	-0.782	0.032	0.537	0.12	?
user2	0.313	-0.046	-0.007	-0.259	0.043	?
….						

From the results, we can observe that by using the following parameters, both the precision and recall of our developed model are higher than that of the random guess.

(1) $\gamma_1 = 2$, $\boldsymbol{A} \cup \boldsymbol{B}$ (Precision: +114%, Recall: +67%)

(2) $\gamma_1 = 10$, only $\boldsymbol{B}$ (Precision: +339%, Recall: +107%)

(3) $\gamma_1 = 10$, $\boldsymbol{A} \cap \boldsymbol{B}$ (Precision: +649%, Recall: +187%)

The performance of $\gamma_1 = 10$ is generally better than that of $\gamma_1 = 2$, since more-related websites are specified as γ_1 increasing. Considering $\gamma_1 = 10$, $\boldsymbol{A} \cap \boldsymbol{B}$ achieves both higher precision and recall than only $\boldsymbol{B}$, probably because the intersection of sets also specifies more-related websites. However, the performance of only $\boldsymbol{A}$ is very worse here, mainly due to the limited number of moved users in the data set.

4. Conclusions

A life event prediction scheme based on the fluctuation of users' web access is proposed in this paper. First, the life event related websites are specified by using two types of statistical difference: the statistical difference of the experienced users during and not during the life events; the statistical difference between the experienced users and the unexperienced users. Especially, the backward method is adopted by us for the first type. Then, the prediction model based on users' fluctuation of accessing these related websites is developed with the logistic regression classifier. The 10-fold cross validation proves that our developed model is able to achieve higher precision and recall simultaneously than the random guess, which may significantly reduce the cost and improve the efficiency of advertising for the industrial applications with a large number of users. The further performance improvement through searching the optimum parameters (γ_1, γ_2, s_{min}) and machine learning model, as well as more evaluation metrics (f1 socre, pr-auc score, etc.) will be considered afterwards.

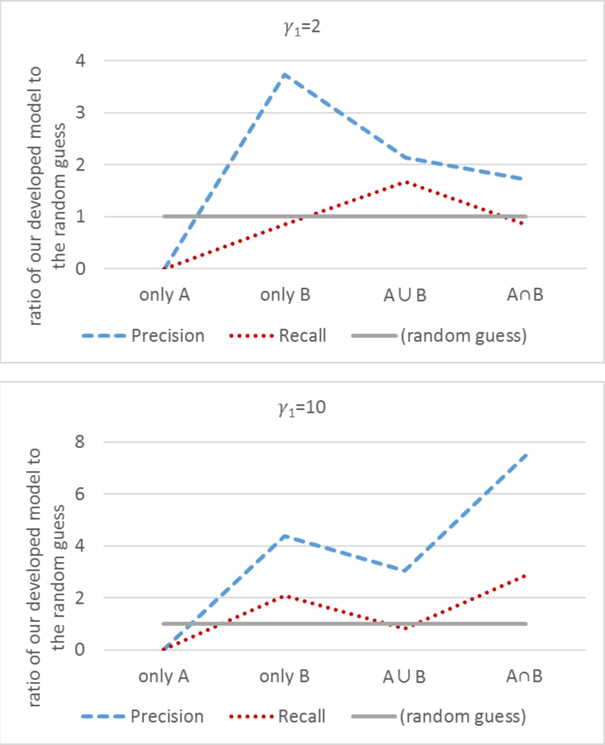

Fig 2. Performance (precision and recall) improvement of our developed model compared to the random guess.

References

[1] Pasquale Lops, Marco de Gemmis and Giovanni Semeraro, "Content-based Recommender Systems: State of the Art and Trends", In Recommender Systems Handbook, pages 73–105. Springer US, 2011.

[2] Xiaoyuan Su and Taghi M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques", Advances in

Artificial Intelligence, vol. 2009, Article ID 421425, 19 pages, 2009. doi:10.1155/2009/421425.

[3] Peng Jiang, Yadong Zhu, Yi Zhang and Quan Yuan, “Life-stage Prediction for Product Recommendation in E-commerce”, Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pages 1879-1888, August 10-13, 2015, Sydney, NSW, Australia.

[4] Bin Guo, Kai Dou and Li Kuang, “Life Stage Based Recommendation in E-commerce”,2016 International Joint Conference on Neural Networks, Pages 3461-3468, July 24-29, 2016, Vancouver, BC, Canada.

[5] Remo Manuel Frey, Runhua Xu and Alexander Ilic, “Mobile App Adoption in Different Life Stages: An Empirical Analysis”, Pervasive and Mobile Computing, vol. 40, Pages 512-527, Sept. 2017.

[6] Shun Abe, Masumi Shirakawa, Takahiro Hara,Kazushi Ikeda and Keiichiro Hoashi, “Construction of Life Event Prediction Model using Tendency of Word Occurrence in User’s Tweet History”,IEICE technical report, vol. 117, no. 108, Pages 1-6, Jun. 2017.

[7] NTT and Tokyo Institute of Technology, unexamined patent application 2011-227746, 2011-11-10.

[8] NTT and Tokyo Institute of Technology, unexamined patent application 2013-125495, 2013-06-24.

[9] Dai Nippon Printing Co., Ltd, unexamined patent application 2017-117351, 2017-06-29.