IADIS International Journal on WWW/Internet
Vol. 12, No. 1, pp. 52-64
ISSN: 1645-7641

USER INTENT PREDICTION FROM ACCESS LOG IN ONLINE SHOP

Hidekazu Yanagimoto. *Osaka Prefecture University. 1-1, Gakuen-cho, Naka-ku, Sakai, Osaka, 599-0011, Japan.*

Tomohiro Koketsu. *Osaka Prefecture University, 1-1, Gakuen-cho, Naka-ku, Sakai, Osaka, 599-0011, Japan.*

ABSTRACT

A lot of recommendation systems on online shops use user's order histories in order to determine recommendation items. In general recommendation systems items are selected based on neighbor users defined according to similarity among users using the order histories. However, the method cannot be applied to new users who have never purchased anything in an online shop and do not define the neighbor users based on their order histories because of undefined similarity. The problem is called a cold start problem. In order to overcome the problem we proposed a method which uses user's access logs to make user profile instead of his/her order histories. Although the access log is less reflective of the user's preference than the order history, we can estimate their intent by careful access log analysis because the access log consists of user's review processes for their purchase. Therefore, to clarify users' intent we use only web pages related to decision of order products strongly. And in order to find these web pages we analyze access logs of users who ordered the same product or the same category. Then we use these pages for making a user profile. In experiments we estimate neighbor users of new users using their user profiles constructed with access logs. And we predict a category of a product which the new users will purchase to examine the efficiency of our proposed method. From experiments we found that there were some categories in which the proposed method can correctly predict new user's target, we confirmed the effectiveness of our proposed method as a solution for cold start problem. However, we found that we improved the proposed method to predict a product itself.

KEYWORDS

Access log analysis, Data mining, Recommendation system, Collaborative filtering, PageRank

1. INTRODUCTION

In online shops, many recommendation systems are used to improve a conversion rate (Kamishima et al 2006). The conversion rate is a purchasing rate for visitors. If we can recommend products that users are requiring or searching, we achieve a high conversion rate. In general recommendation products are selected based on neighbor users defined according to similarity among order histories or users' demographic information registered previously (Linden et al. 2003, Auer et al. 2002 and Gittins et al. 2011). The neighbor users are users who are considered to have a similar preference to a target user. Hence, it is important to find correct neighborhood in the systems. However, there is a problem that we cannot define neighborhood for new users who use the site for the first time or have never purchased any products since we cannot define similarities between the new user and another user. This problem is called a cold start problem (Sahebi 2011, Zhang et al 2010).

In this paper, we overcome the cold start problem using access logs in an online shop instead of order histories since the access logs are obtained more easily. Our aim is to define appropriate neighbor users from access logs and predict new users' intents. However, access logs are not reflective of the user's preference directly, compared with the order history. For example, a user just browses various products according to temporary interests in some cases and compares a few products to decide what to buy according to some concrete interests. So the access logs include some access logs not related to user's purchase as noise. Then, we pay attention to access logs of user's review processes on their purchasing, since the processes sure to have specific intents. And in order to capture user's various intents using web pages we pay attention to common pages in access logs of users who ordered the same product or the same category. This is because that we consider that web pages visited by many customers on their purchasing processes have an effect on the purchasing of the category-specific. Concretely, we construct a network from the access logs. Evaluating the network, we obtain important web pages which are highly relevant to customers' purchase. In this paper, we use PageRank algorithm (Page et al 1998) to select these web pages and call them "*Characteristics Pages*; *CPs*" (Koketsu et al 2012). Using *CPs* in each category, we make a feature vector which elements denote whether the user visited the *CPs* or not as the user's profile. In experiments we estimate neighbor users of new users by calculating similarities among users' profiles. And we predict the category of product which the new users will purchase to examine the efficiency of our proposed method. From experimental results in category prediction we found that there were some categories in which the proposed method can correctly predict the product category that new users actual purchased, we confirmed the positive effectiveness of the proposed method as one solution for the cold start problem.

2. USER INTENT PREDICTION FROM ACCESS LOG

In this section we explain a method to predict user's intents from access logs in online shop. Access logs show processes where users determine whether they should check a web page or not to find a product. Hence, we analyze the access logs and try to select web pages influencing users' purchases. Moreover, defining user profile using the selected web pages, we estimate user intent and predict their purchases from their access logs.

Our proposed method consists of three models to develop user intent prediction algorithm: (1) extract characteristic web pages for purchases from users' access logs related to their purchases strongly in each category, (2) make user profiles based on the extracted web pages, (3) estimate neighborhood users using the user profiles and predict the category related to users' purchase products. We call the selected web pages *CP*s and we construct user profiles using the *CP*s.

2.1 Feature Selection from Access Log

We have to effective *CP*s from access log to capture user intents from access logs. There are many web pages influencing purchases. However, we need to select only web pages having generality to predict other users' purchase using them. Hence, we gather access logs including the same category product purchases and construct a network showing user's transitions before their purchases. We analyze the network using PageRank and select *CP*s to construct user profiles. In this section we explain a *CP* extraction method using PageRank (Koketsu et al. 2012).

First, we construct a network for representing of relationship among web pages using access log as transition information. Especially, we make an a matrix *H* according to the following rule. The matrix *H* denotes the network.

$$h_{uv} = \begin{cases} n_{uv} & (\text{the number of transition } u \to v) \\ 0 & (\text{no transition or } u = v) \end{cases}$$

The matrix *H* shows connections between web pages and frequency of transition from a web page to another web page. In general case used PageRank the matrix *H* is an adjacency matrix and does not include information on transition frequency.

Second, we transfer the matrix *H* to a transition probability matrix *S* for representing of the users' stochastic behaviors. The scale of *H* is a $k \times k$ where the *k* denotes the total number of web pages that all users who buy the same category products visited.

$$s_{uv} = \begin{cases} \dfrac{h_{uv}}{\sum_{v=1}^{k} h_{uv}} (\text{if } h_{u*} > 0) \\ \dfrac{1}{k} (h_{u*}\ is\ zero) \end{cases}$$

After the transformation the following condition holds.

$$\sum_{v=1}^{k} s_{uv} = 1$$

The element of the matrix *S* shows transition probability between the *u*th web page to the *v*th web page based on actual transition frequency. If a web page with no outgoing links exists, the probability that jump to one of the all web pages on the site is given uniformly. That is, the model represents that users can visit any web pages on a network. The model prevents a particular web page monopolizing the value of PageRank. This model is called random surfer model (Page et al 1998, Gleich et al 2010, Langville and Meyer 2009).

We calculate PageRank scores using the transition probability matrix *S*. The PageRank score is defined in an eigenvector corresponding to the maximum eigenvalue of the *S*. Strictly speaking, the score represents probabilities under which user on the network walks according to the *S*. That is, the probability corresponds to a PageRank score as the importance of each web page. Using this idea, we select *CP*s from many web pages included in users' access logs.

Web pages that many customers visited tend to obtain high score. We regard such web pages excluding top pages and order pages and so on as *CPs* which affect users' purchase. Hence, since we use the *CPs* to represent frequently occurring web pages in users' behaviors. The *CPs* is obtained in each category since users take different activities in each category.

2.2 User Profile Construction

Table 1. A sample of user profiles including two categories

	Features (URL)			
Users included in category A and category B	w_1	w_2	$\cdots$	w_{2n}
u_1	1	0		0
u_2	0	1		1
$\cdots$				
u_{P+q}	1	1		1

The *CPs* are extracted in each category using the previous process. Using the *CPs* as features to represent user's intent, we make a user profile form access logs. Table 1 shows an example of user profile. In this case we assume two categories, A and B, the number of users purchasing products in the category A is *p* and ones purchasing products in category B is *q*. In our proposed method we have to decide the number of the *CP*s, *n*, to make user profile previously. If the number of *n* is 100, we use 200 (2*n*) *CPs* for features of the user profile because of two categories. In Table 1 the user profile is a binary vector denoting whether he/she visited *CP*s or not. If a user u_1 visited a *CP* w_1, the value of feature vector in his/her profile is 1. If the user did not visit the page, the value is 0. The more access logs we can use, the many elements in a user profile are occupied with 1.

Under the assumption that a user selects a web pages to discovery necessary products for him/her web page selection reflects his/her intent of purchase. Hence, the user profile includes user's intents because of extraction of characteristic transition from access logs and is effective to predict user's purchase products. Moreover, since we select CPs form all web pages included in access logs, a feature space span with the user profile is a lower dimension space than with all web pages. These characteristics cause less computational costs in defining the neighborhood and avoid a curse of dimensionality.

2.3 User Intent Prediction using User Profile

We use collaborative filtering like approaches to predict user intents using the user profile. In collaborative filtering systems we construct a group similar to a target user and predict the user intents using user intents belonging to the group. Hence, it is important to define the neighborhood of a user. In this study we use user profiles to find the neighborhood.

Using the user profile, we define similarity among users to find similar users. In this study we use cosine similarity to define the similarity between users. The cosine similarity is defined below.

$$c_{ij} = \frac{\sum_{l=1}^{2n} x_{il} x_{jl}}{\sqrt{\sum_{l=1}^{2n} x_{il}^2} \sqrt{\sum_{l=1}^{2n} x_{jl}^2}}$$

The c_{ij} in the equation is the cosine similarity between *i*th user and *j*th user. The x_{il} is the *l*th element in feature vectors of *i*th user. Calculating the similarity among all users according to their user profiles, we obtain a similarity matrix which size is $(p + q) \times (p + q)$. From the matrix we can define the neighborhood user based on the similarity. Using the similarity matrix, we can find users similar to a target user and define the neighborhood of the target user.

To predict user's intent we use the neighborhood in a collaborative filtering fashion. We determine the intents as a majority of neighbor users' intents since the neighbor users take similar transitions. In this study we predict a product category but not products themselves since in extracting *CP*s there are the small number of users buying the same products and *CP*s extracted from small access logs are not reliable.

Our proposed method for category prediction has four steps: (1) we extract *CPs* from existing users' access logs on their purchase processes in each category, (2) using the *CPs*, we define the elements of a profile and construct user profiles for existing users and new users from their visited web pages, (3) When we pay attention to top *N* neighbor users, we determine an estimated category as a category that is occupied highest percentage in the top *N* users, (4) If the estimated category matched the category of the new user, we can predict a correct category (cf. Koketsu et al 2013).

3. EXPERIMENTS

In this section we carry out some experiments using actual access logs in a real online shop to evaluate our proposed method. Since the online shop deals with golf equipment, all access logs and order histories are related to golf.

3.1 Data Set

In experiments we used a dataset containing 1,561,205 access logs and 6,656 order histories in a real EC site. The access log data contains 22 attributes and the order history data contains 12 attributes. From many attributes in dataset we selected some attributes for our method as the following Table 2. These attributes are included in ordinary access logs in a online shop.

Table 2. Attributes included in access logs

Access log data	Order data
・User ID	・User ID
・Order ID	・Order ID
・Time and Date of Event	・Product ID
・Referrer URL	・Order Date
・Access URL	・Category of products
・Product ID	・Kinds of manufacturers
・Session ID	
・Session starting time	
・Session finishing time	

In the order history 44 categories are included, for example, shirt and cap, driver and so on. In this paper, as the target of the experiments we picked up 6 categories since purchase transactions occurred in the categories. We show the number of purchase transactions in the 6 categories in Table3.

Table 3. Order information in each target category

Product's Category	*More than once* (January)	*Once* (February)
Short-Sleeve Shirt	658	149
Driver	1,398	443
Wedge	1,297	227
Visor	1,610	187
Tops	794	73
Cap	527	46

In Table3 *Once* means the number of users who purchase a product in the category only once and we regarded as new user in our experiments. *More than once* means that the number of users who have purchased products more than once. They are candidates of neighbor users for the new user in the site. We regard users who purchase more than once in the site as existing users. And we regard users who purchase a product only once as test data. We use training data to extract *CPs* and define the neighborhood users. (i.e. We obtain a ranking of similarity among a new user and 6,284 existing users each the new user) Finally we measure whether an estimated category matches the category of product that the new user actual ordered.

First, we have to discuss whether *CP*s selected with the proposed method is effective to predict user intents or not. Hence, we focus on discriminative ability of product categories with *CP*s. Speaking concretely, we construct feature vectors of categories using access logs of users who buy the category products. We discuss whether we can separate each category using the feature vectors. In this experiments we use another dataset and show a content of data in Table 4.

Table 4. Categories to check discriminative ability

Product's Category	*Indicator*
Short-Sleeve Shirt	c1
Underwear	c2
Driver	c3
Cap	c4
Wedge	c5

It is desirable that you can approach in an early stage for a new user who visited the site. Therefore in this study, when we make the new user's profile, we changed duration of their access logs. Since we regard a *once* user as the new user in the site. Therefore, we need to consider only review processes and have to reflect them to their profile. We deleted the web pages related to order form from data. Second, we used 30% of the data and increased by 10% in Figure 1.

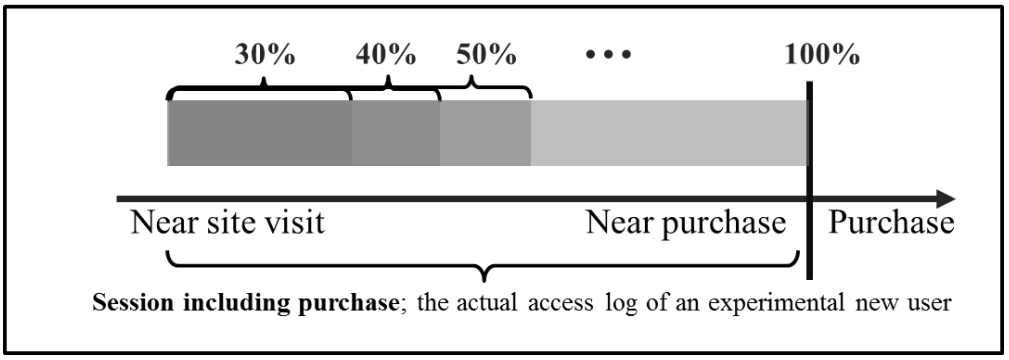

Figure 1. The image of simulation that we change the duration of new user's access log for making his/her user's profile

Second, we discuss how the number of top neighbor users influence precision in category prediction. The too small number of neighbor users causes severe sensitivity to user selections and the too large number of neighbor users causes noisy prediction. We observe the precision of category prediction according to the number of neighbor users.

Third, we verified how the number of *CP*s influences precision in category prediction. The too small number of *CP*s is less discriminative. Hence, we check the appropriate number of *CP*s using actual access logs.

3.2 Results

3.2.1 Discriminative Ability according to *CP*s Selection

The discriminative ability depends on *CP*s selection since a user profile is described with visited *CP*s. Hence, we discuss similarities among 5 categories. In Figure 1 varying the number of CPs, we measured similarities between category 1 and 5 categories. We found user profiles with *CP*s could separate category 1 and other categories. However, using all web pages as *CP*s, the similarities are the same values and they could not discriminate categories.

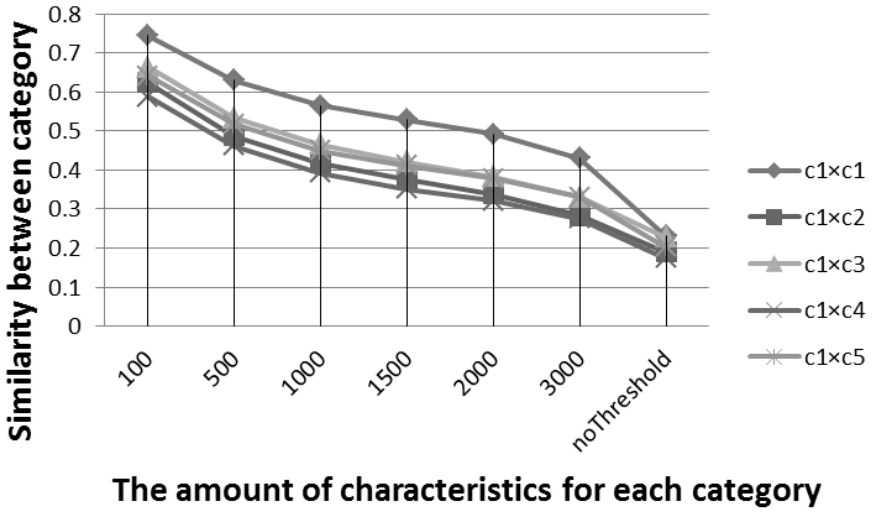

Figure 2. Similarities among 5 categories

3.2.2 Prediction Precision according to the Number of Neighbor Users

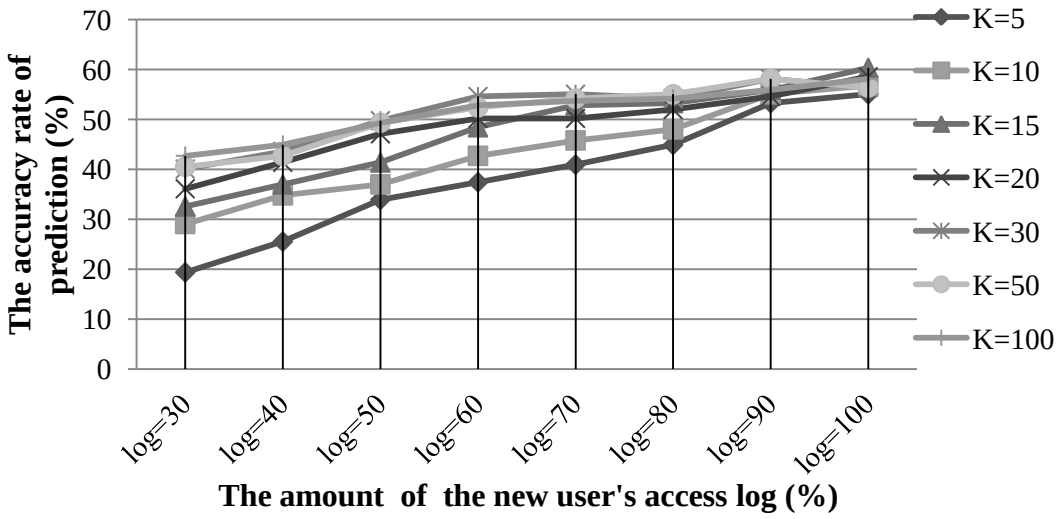

Figure 3. The result of category prediction of 227 new users in wedge category (we focus on the number of neighborhood users *K* in category prediction)

Fig.3 shows that the accuracy rate of category prediction increases when the amount of new user's logs increase. We consider that this is because the user's intent related to purchase would be so clearer in just before their purchasing that their preference is reflected accurately in the profile. From Fig.3, we found that the result of K = 100 is the most accurate. It is

concerned that we can make predictions accurately by the effect of a vote from a lot of neighborhood users. However, the accuracy rate is converged to around 60% regardless of the amount of access log is also increasing close to just before purchasing. Therefore, it is necessary to consider the value of *K* and the additional recommendation method.

3.2.3 The Amount of *CPs* used for making the User's Profile

Then, we consider the relationship between the accuracy rate and the amount of CPs for making user's profile. Fig.4 is the result of prediction of 443 new users in driver category.

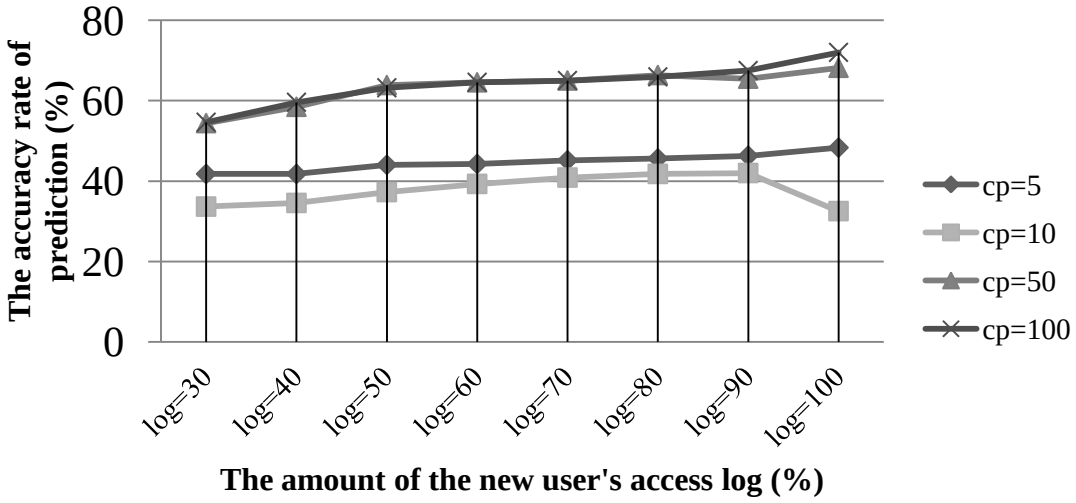

Figure 4. The result of category prediction of 443 new users in driver category (we focus on the amount of *CPs* for making the user's profile)

Fig.4 shows that the accuracy rate of category prediction is increased gradually when the amount of new user's logs increased. In the driver category, even when the amount of log is 30% of new user's log, we obtain at least 30% or more accuracy rate. In other words, though it is difficult to know 443 new users' intents in conventional methods, it is possible to realize the correct approach to 132 new users in proposed method. The conventional method cannot define neighborhood users. That is, these systems cannot personalize services. So they cannot recommend personalized products but select products as like a ranking of hot-selling products or categories. Here, we show the information about the hot-selling category in January from the same experimental data.

Table 5. The hot-selling category ranking information in January

Rank	*Category*	*Order amount*
1	Ball	9,886
2	Gloves	5,117
3	Outer (Men's)	4,953
4	Tee	3,709
5	Driver	3,690
6	Wedge	3,278
12	Cap (Men's)	2,600
18	Tops (Men's)	1,190
21	Short-Sleeve Shirt (Men's)	1,091
25	Visor (Men's)	817

As shown in Table.5, if we use the ranking information which is not considered user's preference, the accuracy rate becomes 0% in terms of the category estimation. Therefore, the proposed method might be more effective than the conventional method as like ranking information.

In addition, we found that as parameters were *log=90~100* and *CP=5*, the accuracy rate fell. That is, the result shows that new user's log increase is not always to improve the predictive accuracy.

3.2.4 Comparison among Results of Experimental Categories

In this section, we explain the differences among the results of the categories. We show the result for all categories in Fig.5. The number of neighborhood users *K* is 100, the amount of *CPs* for making user's profile is 100.

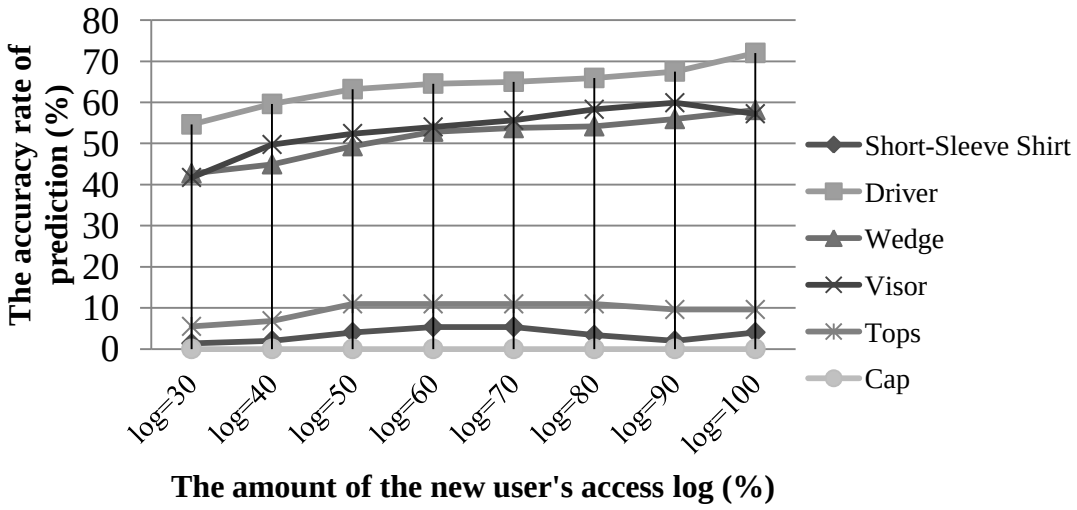

Figure 5. The result of category prediction for all categories (the number of neighborhood users *K* = 100 for prediction, the amount of *CPs* = 100)

From Fig.5, in Driver and Wedge, Visor categories we obtained a high accuracy rate. On the other hand, in other categories we obtained only 1% to 10% accuracy rate regardless of the amount of the new user's log. We consider main reason is lack of data. For, as you can see in Table.3, the number of customers in the categories which have low accuracy rate is less number of orders than other categories. That is, since the customers' access log for extracting the *CP*s is small, the characteristics of customers may not be reflected to *CP*s. Therefore, it is necessary to consider the optimization for the amount of CPs for prediction and how to select the CPs.

3.2.5 Discussion of CPs' Types

In Figure 5 prediction precision vary depending on categories. Hence, we discuss types of *CP*s using their directory structure. In this case we focus on "Product", which denotes web pages related to products, and "item", which denotes web pages related to more concrete products. Generally "Product" is a wider category than "item".

Table 6 shows *CP*s in high precision categories including more "item" than low precision categories. The result shows in categories where we can predict with high precision *CP*s included more detailed category. We have to discuss relationship between types of *CP*s and prediction precision.

Table 6. Directory name of URL in *CP*s

Category	*"Product" included in URL*	*"item" included in URL*
Short-Sleeve Shirt	12	39
Driver	12	112
Wedge	15	99
Visor	15	106
Tops	12	62
Cap	7	45

3.2.6 Product Prediction using User Profile

Finally we predict purchase product but not purchase product category. Figure 6 shows product prediction precision in Driver category. In this experiments we determined recommended products which the neighbor users bought. The precision is very low although the number of neighbor users varies. To improve the precision in product prediction we have to discuss prediction approaches again. For example, in prediction mixing some prediction algorithms, we try to improve precision.

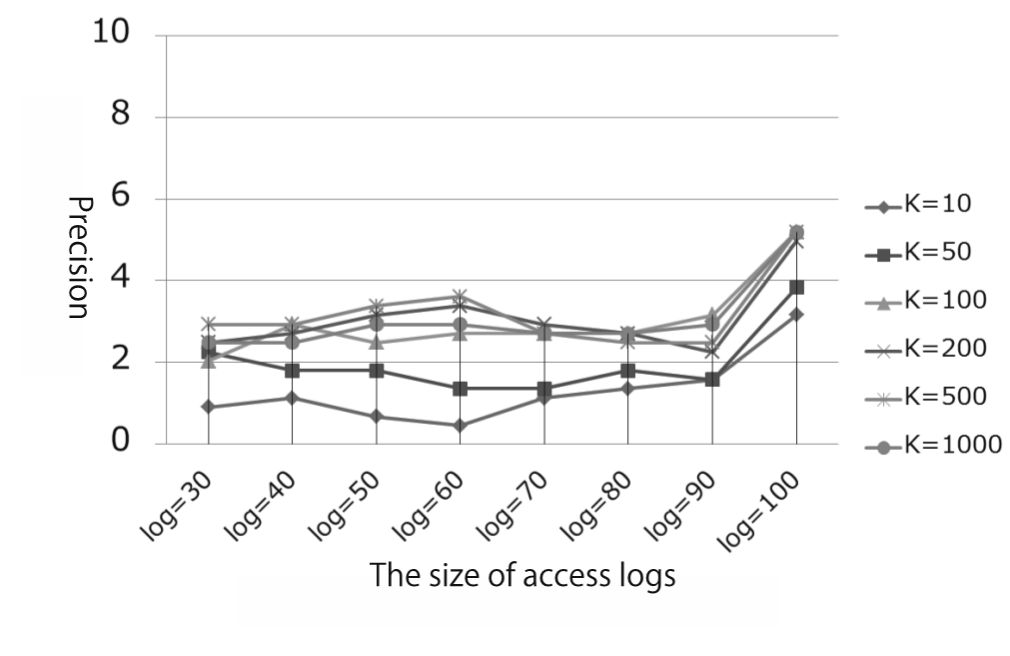

Figure 6. Product prediction in Driver category using user profil

4. CONCLUSION

In this paper, we conducted the neighbor user estimation from users' access logs on their purchase processes. Furthermore, in order to confirm the effectiveness of our proposed method as a service we predict new users' purchasing product categories. From the results we confirmed that our proposed method could predict the new user's purchasing categories in early stages. Hence, our proposed method is effective against cold start problem since conventional method cannot solve it. However, our proposed method could not predict a product itself effectively. To improve it we use more access logs to select *CP*s and construct improved user profile.

We have to improve the accuracy of prediction for operating on real service and it is desirable that you can develop efficient product recommendation not category. To achieve it we need to conduct experiments with various experimental conditions and confirm the validity of *CPs* extracted from access logs as users' intents and the optimal parameters.

ACKNOWLEDGEMENT

We would like to thank Golf Digest Online Inc. for giving us online shop's data.

REFERENCES

Linden G., Smith B., and York J., 2003, Amazon.com Recommendations: Item-to-Item Collaborative Filtering, IEEE Internet Computing, Vol.7, No. 1, pp.76-80.

Gittins J., Glazebrook K., and Weber R., 2011, Multi-armed Bandit Allocation Indices 2nd Edition, John Wiley & Sons Ltd, UK.

Auer P., Cesa-Bianchi N., and Fischer P., 2002, Finite-time Analysis of the Multiarmed Bandit Problem, Machine Learning, Vol. 47, pp.235-256.

Koketsu T., Yanagimoto H., and Yoshioka M., 2012, Access Log Analysis for EC Service with PageRank, *Proceedings of The First Asian Conference on Information Systems (ACIS 2012)*, Siem Reap, Cambodia, pp.80-85.

Koketsu T., Yanagimoto H., and Yoshioka H., 2013, Neighborhood User Estimation from Web Access Log in EC Service, *4th International Conference on E-Service and Knowledge Management (ESKM 2013)*, Matsue, Japan, pp.89-94

Koketsu T., Yanagimoto H., and Yoshioka M., 2013, Access Log Analysis with PageRank, *IEEJ Transactions on Electronics, Information and Systems* Vol.133 No.7 pp.1–6.

Page L., Brin S., Motwani R., and Winograd T., 1998, The PageRank citation ranking : bringing order to the Web, Technical report, Stanford University.

Gleich D., Constantine P., Flaxman A., and Gunawardana A., 2010, Tracking the Random Surfer: Empirically Measured Teleportation Parameters in PageRank, In *Proceedings of the* 19*th international conference on World wide web* (*WWW*2010), Raleigh, USA, pp. 381-390.

Langville N. and Meyer D., 2006, Google's PageRank and Beyond, Princeton University Press, USA.

Sahebi S. and Cohen W., 2011, Community-Based Recommendations: a Solution to the Cold Start Problem, *Workshop on Recommender Systems and the Social Web (RSWEB) held in conjunction with ACM RecSys'11*.

Kamishima T., and Akaho S., 2006, Nantonac Collaborative Filtering — Recommendation Based on Multiple Order Responses, *The 1st International Workshop on Data-Mining and Statistical Science (DMSS)*, pp.117-124.

Kamishima T. and Akaho S., 2010, Nantonac Collaborative Filtering — A Model-Based Approach, *The 4th ACM Conference on Recommender Systems (RecSys)*, pp.273 -276.

Zhang Z., Liu C., Zhang Y., and Zhou T., 2010, Solving the cold-start problem in recommender systems with social tags, ELP