Improving Advertisement Recommendation by Enriching User Browser Cookie Attributes

Liang Wang, Kuang-chih Lee, Quan Lu
Yahoo! Inc.
701 First Avenue
Sunnyvale, CA, US
{wlliang@, kclee@, qlu@}yahoo-inc.com

ABSTRACT

User attributes including online behavior history and demographic information are the keys to decide whether a user is the right audience for an advertisement. When a user visits a website, the website generally plugs a browser cookie string ('bcookie' for short). The bcookie is then used as an identifier to collect the user's online behavior, as well as the joint key to link user profile attributes, such as demographic information and browsing history. However, the same users can have different bcookies across different browsers and devices. Moreover, bcookies can expire after some period, be cleared by browsers or users. This situation of bcookie discounting typically introduces both performance and delivery problems in online advertising since advertisers are hard to find the most receptive audiences based on the user profile information. In this paper, we try to tackle this problem by using an 'assistant identifier' to find the linkage between different bcookies. For most of the Internet company, in addition to the bcookie information, there are always other identifiers such as IP address, user agent, OS type and version , *etc.*, stored in the serving log data. Therefore, we propose an unified framework to link different bcookies from the same users according to those assistant identifiers. Specifically, our proposed method first constructs a bipartite graph with linkages between the assistant identifiers and the bcookies. Next all attributes associated with each bcookie are propagated along the graph using the state-of-the-art random walk model. Offline comparative experimental studies are conducted to confirm that by enriching the bcookie attributes we can recover 20% more online users whose bcookie information is lost, which is greatly helpful to delivery more budget spending with a little loss in precision of predicting converted users. On-product evaluation further confirms the effectiveness of the proposed method.

Keywords

Advertisement recommendation; user attribute enrichment; probability mass propagation

1. INTRODUCTION

After many years of exploration, online advertising has become more mature with very large market shares. In 2014, spendings on online advertising approached more than 50 billion dollars in the whole world [5]. Different from traditional media advertising like TV and newspaper, the online advertising systems know more about the users' attributes including their demography information and propensities in products, which are noted by some registration information and user online behaviors. Thus the online advertising is capable of serving more proper advertisement to the users, and meanwhile yields more returns for the advertiser by the predictable user response. Therefore the key to the success of online advertising is the completeness of the user attributes.

To simultaneously protect the privacy of users and collect the users' attributes from their registration information and behavior history, the online users' browsers are assigned a browser cookie ('bcookie' for short) by a website when they first visit this website. The website then track the users's activity until the bcookies expires or deleted by the users. However, as studied in literature [3][8], users may frequently clear their cookies (so called 'cookie churn' problem), and one user may associate with multiple bcookies when they are working with multiple browsers or multiple devices. As a result, most of the bcookies we are dealing with typically have fragmented user attributes or even no user attribute at all.

To provide a better personalized advertisement serving experience, one essential step is to identify the underlying real user behind the bcookies so that we can link the fragmented user information among a set of bcookies together to form a complete user profile. In [8], the author provided a thorough study of whether using a single user identifier (including IP address, user agent, bcookie) or the combination of user identifiers can correctly identify a user client host. The study shows that bcookie is the best user uniqueness identifier except the registrated user ids and the major reason for the bcookie mismatch is the cookie churn problem. However, this task is difficult due to the lack of real user level identification information.

In the literature, there are several studies [3][2][9] that focus on linking the bcookies together according to whether they belong to the same browser or device. In [3], the authors propose a graph coloring method to cluster the bcookies with no lifetime overlap (since they belong to the same browser) and similar in their behavior patterns including visiting webpage categories and IP used. They use the Yahoo toolbar id as the ground truth to measure the performance

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA
© 2016 ACM. ISBN 978-1-4503-4073-1/16/10. . . $15.00
DOI: http://dx.doi.org/10.1145/2983323.2983374

of their method. The method has limited applications due to the fact that all bcookies form the same users needs to be stored in the same browsers; however, since users may switch to different browsers or different devices frequently nowadays, this method is less applicable to mobile related advertising business. Yen *et al.* in [9] studied the problem of assigning the search queries to multiple users working on a single machine. Comparing to the bcookie grouping task, this problem assumes there are fixed machine identifiers so the problem is simplified. In summary, most of the existing approaches only concern the correctness of the bcookie grouping (into one host client or browser), and less studies have been made on whether clustering the users to enrich the bcookie attributes really improves the performance of applications like advertisement recommendation.

In this paper, we propose the solution to a more general problem of enriching bcookie attributes from a set of the bcookies which are likely to arise from the same underlying user. Based on the enriched user attributes, we also study the performance improvements on the advertisement recommendation problem. Our study is conducted on Yahoo advertising serving logs. For each raw log, the user is not only identified by bcookie, but also other identifiers like IP address, user agent, device type, *etc.* We build a bipartite graph by linking the bcookies to another identifier (let's call it 'assistant identifier' in the following sections) according to the logs, and enrich the attributes associated with each bcookie by the propagation method that executes the state-of-the-art random walk algorithm with an iterative updated transition probability on the graph, so that the bcookies can transmit their attributes to the ones which are strongly linked to them by the assistant identifier. The intuition behind this method is that bcookies that shares similar assistant identifiers are likely to be raised from the same underlying user or the closely connected group of persons (like the family members). Since we are dealing with large scale data with billions of bcookies every week, the proposed method is designed to support parallel computing. We also study the performance of the proposed enrichment method by the application of finding potential customers in advertisement recommendation. Experimental results confirm that, by enriching the attributes of the bcookies, we are capable of discovering more potential customers for online advertising even for bcookies without any history.

1.1 Organization

The rest of this paper is organized as follows. The proposed algorithm to enrich the bcookies' attributes are detailed in Section 2. The performance study of enriching the bcookies' attributes for finding potential customers in advertisement recommendation is presented in Section 3. We conclude the proposed method in Section 4.

2. ENRICH THE BCOOKIES' ATTRIBUTES BY ATTRIBUTE PROPAGATION

To enrich bcookies' attributes by graph propagation between 'linked' bcookies, we need to solve two key issues: (1) find the linkage and the strength of the linkage between different users; (2) since we are working on large scale internet company level data, the computational algorithm should be scalable well. Specifically, we conduct our study on Yahoo's display advertising platform.

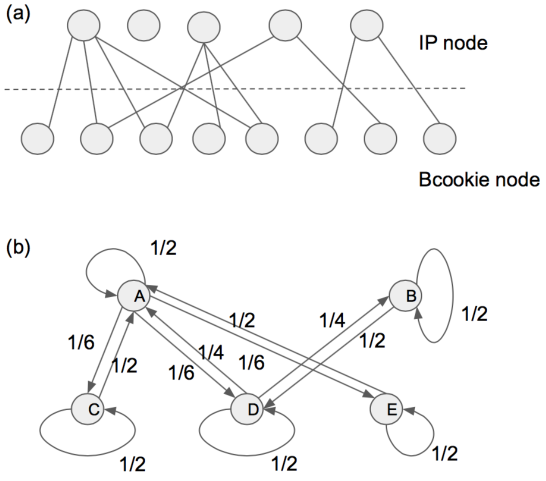

Figure 1: (a) Illustration of building the linkage between bcookies by the assistant identifier (IP address), when a bcookie and an IP appeared together in a log, we link them by an edge; (b) demonstration of the method to compute the transition probability between nodes for random walk.

2.1 Find linkage between bcookies

In the advertising logs, besides the bcookie, the user in each serving log is also indexed by other user identifiers including the IP address, user agent, device type *etc.* Although the bcookies themselves are independent to each other, some of them may have the same IP addresses, user agents, *etc.* We use these identifiers as 'assistant identifiers' to link different bcookies. According to our data analysis, among these assistant identifiers, the IP address is the most informative one compare to others and is more robust with the smallest missing rate in logs. In this paper, we carry out our study by using IP address as assistant identifier.

As illustrated in Fig. 1 (a), we build a bipartite graph to represent the connection between IPs and bcookies. A link between an IP node and a bcookie node is formed only when they appeared in one serving log together. In reality, there are some commercial IPs shared thousands of users, and there are also fraudulent bcookies fake their traffic as coming from hundreds of IPs. These connections will bring noises to the graph. Hence, we deploy the following prefiltering strategies before propagation to remove the suspicious bcookie-IP pairs.

1. Abnormal impression volume rule: filter out bcookies with number of impressions larger than $threshold_1$
2. Abnormal IP address association pattern: filter out bcookies associated with more than $threshold_2$ of IPs
3. Commercial IP address rule: filter out IPs associated with more than $threshold_3$ of bcookies

In our experiment, we analyze the data for one month and set the thresholds as $threshold_1 = 10000$, $threshold_2 = 50$ and $threshold_3 = 100$. These rules may filter out the significantly suspicious bcookie-IP pairs but leave the nonsignificant ones still in the graph. As we will present in the following sections, we use edge weighting method to minimize the impact of these noisy pairs.

2.2 Propagate bcookie attributes by random walk

We propose the following two assumptions when designing the propagation algorithm on the bipartite graph: (1) when two bcookie nodes shares many IP addresses, they are likely to belong to the same underlying users or a closely connected groups of people (like family members); (2) the more outer links a node (IP or bcookie) has, the more likely the node would be some fraudulent IPs/bcookies or commercial IPs, and we are less confident in transmitting their attributes to their linkagees. Regarding to criterion (1), we define the transition probability between nodes according to markov chain formulation [4]. Starting from a node v_s, the probability of arriving at v_e in t step is:

$$P^t(v_b, v_e) = \begin{cases} w(v_b, v_e) & \text{if } t = 1 \\ \sum_k w(v_k, v_e) \cdot P^{t-1}(v_b, v_k) & \text{if } t > 1 \end{cases} \quad (1)$$

where $w(v_i, v_j)$ represents the transition probability from v_i to v_j, and is equal to 0 when there is no edge between them. By defining the transition probability this way, when two bcookies share more IP nodes in common, they transmit their attributes to each other with higher probability.

Considering the criterion (2), we design the edge weight (transition probability) from node u to v in the bi-direct graph according to the random walk model [6], *i.e.*

$$w_{u,v} = \begin{cases} 1/2 & \text{if } u = v \\ 1/(2 * d(u)) & \text{if u and v are linked} \end{cases} \quad (2)$$

where $d(u)$ is the outer degree of node u. In this way, the edge weights between nodes are asymmetrically defined and the more outer links a node has, the less confident it can transmit its attributes out to other nodes. An example to compute the transition probability between nodes is shown in Fig. 1 (b). For the node A, its out-degree is 3, such that its transition probability to each outer links (to node C, D and E) are 1/6, and 1/2 to itself.

According to this design, for each attribute associated with a bcookie, we set its initial weight to be 1, then, we transmit the attribute with the transition probability defined in the graph and set its weight in the target bcookie (or IP) according to Eq. 1. The final attribute set associated with a node can be determined by thresholding the weights of these attributes.

Since the markov chain propagation process can be performed by matrix multiplication [4], and in our case, most items in the transition matrix defined on Eq. 2 is 0. We employ the distributed sparse matrix multiplication method in [1] to compute propagation probability between nodes as Eq. 1 to compute the final attributes associated with each nodes and their weights. There is a debate between the number of iterations we should perform. On the one hand, when we run more iterations of attribute propagations, we can have more information for the bcookies, however, these informations is more likely to contain noise; on the other hand, when we iterate less, we are more confident in the final bcookie attributes, but we will still face the lacking of information phenomenon. In Fig. 2, we show the result of the enrichment performance changes when using different numbers of iterations. It can be seen, after 3 iterations, the number of bcookies being enriched does not increase too much under more iterations. Hence, we only perform 3 iterations as our experimental protocol.

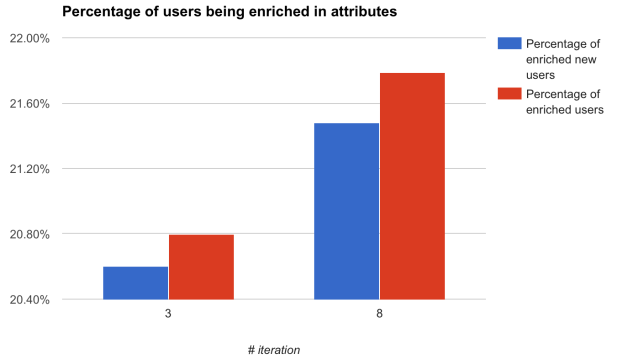

Figure 2: Illustration of the enrichment performance changes using different number of iterations.

Table 1: Performance in bcookie attributes enrichment. Here, 'new' means the user has no attributes in our logs and 'old' means the users are associated with certain attributes

Method	percentage of new users being enriched	percentage of old users being enriched
IP grouping	20.5%	19.7%
Propagation	20.6%	20.8%

3. EXPERIMENTS

We conduct extensive offline and online evaluation of the proposed method on Yahoo's display advertising platform which serves billions of bid request per-day. Since most of the existing works [3][2][9] try to group the users according to their underlying real-world users, we use the user grouping method as a benchmark for comparison purpose. The benchmark method proceeds as follows: first, we apply the same pre-filtering rules as in Section 2 to remove the fraudulent/commercial IP addresses and bcookies from the data. Then, we compute the number of impressions (the number of log items) $N_{bcookie,IP}$ for each $< bcookie, IP >$ pair, and assign the bcookie to the IP with the largest number of $N_{bcookie,IP}$. Finally, the attributes sets of all the bcookies in the same IP group are shared to each other to enrich their attribute set.

Nowadays, the advertisers begin to concern more about the conversion rate of their advertisement, *i.e.*, the number of bcookies found by our system can finally purchase the products related to the recommended advertisements. In this paper, we use the purchase history as the attributes for bcookies to study the performance changes. However, the proposed method can be applied to any attributes for bcookies.

We collect one month's advertisement serving log data to perform the attribute enrichment experiment. There are more than one billion of unique bcookies appeared in the logs. The enrichment impact is illustrated in Table 1. We can see that both the propagation method and the grouping method enriched about 20% of bcookies previously recorded with no purchase history, this is a large contribution to solve

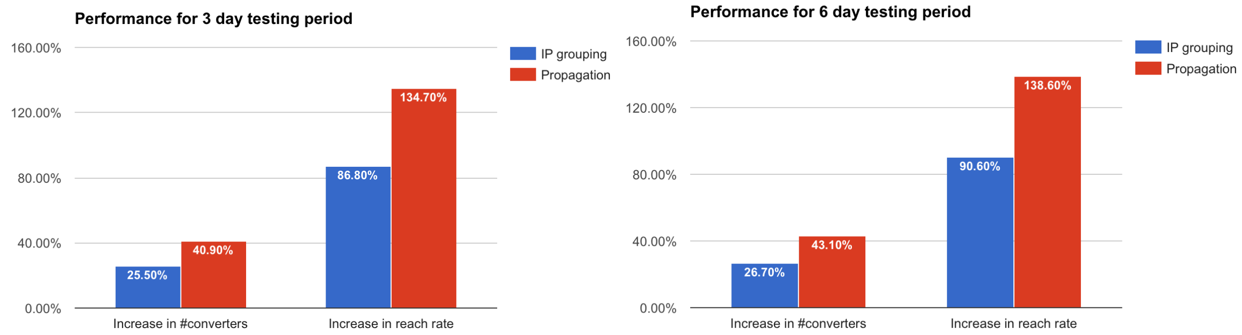

Figure 3: Illustration of the performance changes in advertisement recommendation when applying the enrichment methods.

Table 2: On production performance.

Increase in revenue	Increase in conversion rate
1.86%	1.92%

the cold start problem of our system. Since we are more interested in the quality of the enrichment process, we conduct experiments on advertisement recommendation performance improvements. Specifically, we focused on the purchase prediction improvement brought by the enrichment methods. We use the item-based collaborative filtering algorithm [7] to predict the potential purchased products of the bcookies and recommend them with the advertisements related to these products. In the experiment, we select the top 50 recommendations with score larger than 0 from the output the collaborative filtering algorithm.

We use the enrichment results above, and the following M ($M = 3$, 6) days' bcookie advertisement serving history and purchase history as the ground truth to evaluate the performance. Fig 3 provides the performance changes brought by the attribute enrichment methods. As can be observed from the result, the propagation based method is capable of discovering more converters (the bcookies purchase the products related to recommended advertisements) and reaching more online users than the grouping based approaches. Since in the online serving, there are competition between different advertisements and different advertisement bidding systems, larger reach rate gives the system more chance to win the recommended bcookies for certain advertisements.

3.1 On production test

We put the propagation based enrichment method for online test on two campaigns for two days. The online system runs by combining a set of recommendation modules in a comparative manner. To simplify the implementation, we directly recommend an IP list for each campaign associated with the recommended bcookies. The results are presented in Tab. 2. Compare with the performance of the previous two days (the same weekdays as the testing days), we see the improvement in both the revenue (which is related to the number of advertisements impressions we served) and the conversion rate.

4. CONCLUSION

In this paper, we try to solve the fundamental problem of cookie loss, which leads to the result that user information becomes fragment. This problem is important because it introduces significant performance and delivery drop for online advertising business. In this paper, we propose to link the bcookies through the 'assistant identifiers' which are logged together with the bcookies in advertising serving logs. The attributes of the bcookies are then enriched based on the linkage. Instead of finding ground truth data for validating the correctness of the attribute enrichment, we evaluate both the online and offline advertising recommendation to proof the effectiveness of the proposed methods. As future work, we would like to incorporate more user attributes to improve the performance of our proposed method.

5. REFERENCES

[1] A. Buluc, J. Fineman, M. Frigo, J. Gilbert, and C. Leiserson. Parallel sparse matrix-vector and matrix-transpose-vector multiplication using compressed sparse blocks. In *Symposium on Parallelism in algorithms and architectures*, pages 233–244, 2009.

[2] D. Coey and M. Bailey. Peopel and cookies: Imperfect treatment assignment in online experiments. In *to appear in WWW'16*, 2016.

[3] A. Dasgupta, M. Gurevich, L. Zhang, B. Tseng, and A. Thomas. Overcoming browser cookie churn with clustering. In *WWW'12*, pages 83–92, 2012.

[4] F. Fouss, A. Pirotte, J. Renders, and M. Saerens. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transaction KDE*, 19(3):355–369, 2007.

[5] K. Olmstead and K. Lu. Digital news - revenue: Fact sheet. *State of The Media*, 2015.

[6] K. Pearson. The problem of the random walk. *Nature*, 72(294), 1905.

[7] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *WWW'01*, pages 285–295, 2001.

[8] R. White, A. Hassan, A. Singla, and E. Horvitz. From devices to people: Attribution of search activity in multi-user settings. In *WWW'14*, pages 431–442, 2014.

[9] T. Yen, Y. Xie, F. Yu, R. Yu, and M. Abadi. Host fingerprinting and tracking on the web: Privacy and security implications. In *NDSS'12*, 2012.