December 28, 2018

User Identification Across Online Providers

Anonymous

User Identification Across Online Providers

ABSTRACT

This disclosure describes construction of a user identification model based on confirmed user data obtained from different providers. The user identification model is usable by individual providers to provide missing pieces of user-specific data. An identification server maintains user data that includes globally unique user identifiers. The user data can be utilized to match user actions across different online providers and to perform analyses based on matched actions.

KEYWORDS

user identification; user profile; online tracking; ad conversion; profile matching

BACKGROUND

Users access various online providers, e.g., via websites, mobile apps, etc. to perform various activities. Activities such as viewing text, watching multimedia content, shopping, connecting with friends via social networks and messaging platforms, creating and editing documents, etc. To perform such activities, users may utilize different devices such as smartphones, computers, tablets, wearable devices, etc.

The different online providers can obtain different pieces of user information provided by the user such as user's name, email address, phone number, physical address, birthdate, membership identifier, etc. For example, such information is obtained by providers that require user registration but not by other providers. Providers obtain information about user actions from browser-based parameters such as cookies, history of files downloaded, time spent at the content provider's website, etc. Providers also obtain other information such as a device identifier and other device-specific information (e.g., hardware identifier, operating system/browser, device type, etc.), IP address, sensor data from the device, etc.

Because each online provider obtains only such information as available from the user, each provider has incomplete information about the user. For example, a mobile app provider can obtain a device identifier of the user device, but not any browser-specific information. A shopping website can obtain the user's name and contact information, but not a device identifier. A fitness app can obtain the user's age, but not the user's name; etc. Analysis of user actions is prone to error or cannot be performed with such incomplete data because it is difficult to accurately match user information across the different online providers.

DESCRIPTION

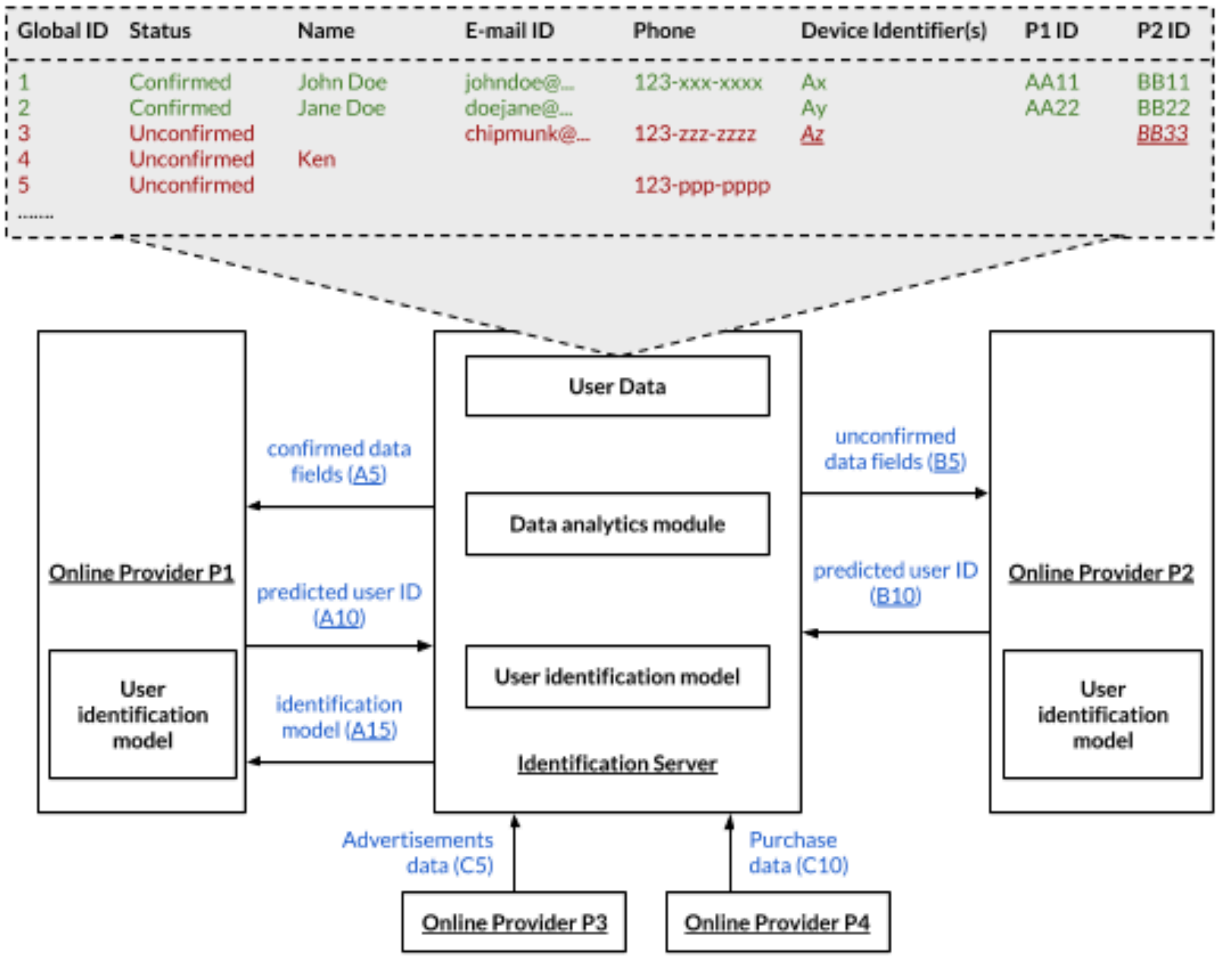

Fig. 1: User identification and analytics across multiple online providers

Fig. 1 illustrates an example of performing user identification and analytics across multiple online providers. An identification server is provided in communication with different online providers (e.g., P1 and P2). For example, the identification server is operated by a third-party distinct from the online providers. The identification server stores user data provided by individual online providers while restricting access to such data with other online providers.

As shown in Fig. 1, user data stored on the identification server generates confirmed data (with status "confirmed") about some users (users with global ID 1 and 2) by utilizing a variety of data sources, including data from online providers. Confirmed data includes data provided directly by users, e.g., during interaction with a digital content provider or service provider, while making an online purchase, during identity verification, etc. Or data associated with a high accuracy level, e.g., based on other user information such as cookies, login names, payment-related information, etc.

User data stored on the identification server also include unconfirmed data about some users (users with ID 3, 4, and 5), e.g., data associated with a low accuracy level. The user identifiers maintained by the identification server are global in nature, e.g., such that no two confirmed users are assigned the same global ID.

User data also includes provider-specific user identifiers, e.g., P1 ID and P2 ID, as illustrated in Fig. 1. For confirmed data, the provider-specific user identifiers are known, for at least a subset of providers. For unconfirmed data, provider-specific IDs may be unknown (users with global ID 4 and 5) or partially known (user with global ID 3).

Generating user identification model

The identification server (or a separate server) also includes a user identification model. The user identification model is generated as follows (steps A5-A15 shown in Fig. 1):

1. The server sends different combinations of fields of data corresponding to confirmed users to different online providers. For example, for the confirmed data shown in Fig. 1, such combinations can include (email+phone); (phone); or other combinations of fields. Different (partial) combinations are provided to different providers. The server requests the online provider to reply with data fields for the user, excluding data fields sent to the online provider - e.g., a particular data field such as device identifier, or any data field available to the online provider.
2. Online providers match the information received from the server with available user information and send a reply with a predicted user identifier, e.g., the device identifier. The server determines a prediction accuracy score for each online provider. Provider P1 receives the information ("johndoe@..." and "123-xxx-xxxx") and returns a device identifier "Ax" for the user, the prediction accuracy score is high, but if it returns an incorrect device identifier, the prediction accuracy score is low. While the figure illustrates a small number of fields, during actual use, a large number of data fields (e.g., 10, 20) can be used to determine the prediction accuracy score. A provider that gets a higher proportion of predictions correct is assigned a high prediction accuracy score. Prediction accuracy scores are specific to the combination of information provided and the prediction made by the provider.
3. Based on the prediction accuracy score of different online providers, the identification server generates a user identification model. The user identification model (e.g., a machine-learning model, a statistical model, etc.) can be utilized to match partial

information in the user data (e.g., unconfirmed data for user IDs 3-5) with information available at different providers to predict the user ID, as explained below. The user identification model can be provided to the online providers.

The user identification model can be, e.g., a machine-learning model, such as a neural network (e.g., a convolutional neural network), Bayesian networks, regression model, decision tree, inductive logic programming, support vector machine, principal component analysis, or a combination of the foregoing. Training data for the machine-learning model includes data fields from user data with confirmed status and excludes other fields. The model is trained to predict the user identity, e.g., of the excluded fields, and/or to determine a data source (e.g., a particular content provider) that is associated with high accuracy. Predictions generated by the model are compared with known data to determine a loss function (a prediction accuracy of the model). The loss function is utilized to adjust features of the machine learning model, such as weights, parameters, or prediction accuracy scores utilized for the prediction. The process of providing training data, generating predictions using the model, and evaluation and feedback using a loss function can be repeated until the model achieves satisfactory level of accuracy.

Utilizing user identification model to match users across different providers

The user information model is utilized to identify users as follows (steps B5-B10 in Fig. 1):

1. The identification server sends different data fields for users that are unconfirmed to various online providers. For example, the server sends the e-mail ID, and/or phone (“chipmunk@...” and “123-zzz-zzzz”) for global ID 3 and requests available information (e.g., device ID and provider-specific user ID) from the online providers.

2. Different online providers match the received data fields with information available to that online provider and respond to the identification server with predicted user information. For global ID3, provider P2 determines a device identifier ("Az") and a provider-specific ID ("BB33") by matching the received e-mail ID and/or phone with available information. The online provider or the identification server utilize the user identification model to associate prediction accuracy scores with the predicted user information. A higher score is associated with providers that have previously provided accurate device identifiers based on e-mail ID and phone. Predictions obtained from different providers are evaluated based on their respective accuracy scores to obtain resolved information. The resolved information (e.g., device identifier "Az") is added to the user data stored by the identity server along with the provider-specific identifier. When the resolved information reaches a high accuracy level, the status of the particular global ID is updated to confirmed.

Performing analytics using data received from providers

An online provider (P3 in Fig. 1) can send data to be analyzed by a data analytics module of the identification server. A website that displays advertisements can send advertisements data (c5) that includes provider-specific IDs for a first set of users that viewed an advertisement for a particular item, e.g., item I, to the identification server. Other providers, e.g., an e-commerce provider (P4), can send purchase data (C10) that includes provider-specific IDs for a second set of users that purchased the item I. The identification server appends such data to the user data, e.g., by adding columns for advertisements and purchases (not shown).

The identification server generates analyses based on such data aggregated from different online providers. The identification server can determine a proportion of users that had viewed

an advertisement for item I on a website provided by online provider P3 prior to purchasing the item I via e-commerce provider P4. Results of the analyses can be sent to the different providers. The results may be sent to an advertiser such that the advertiser can determine the impact of advertisements displayed via different providers.

The availability of a user identification model with the use of model generation as described herein enables different content providers to receive partial user data and match it with locally available data. User data maintained by the identification server is enhanced based on the determined local matches. Such user data include global IDs that are specific to individual users and allows matching user actions across different providers without revealing data from individual providers to other providers. Analytics on user actions can then be performed in aggregate and results be provided to different providers. Use of user-specific identifiers derived from information from different providers provides improved accuracy.

CONCLUSION

This disclosure describes construction of a user identification model based on confirmed user data obtained from different providers. The user identification model is usable by individual providers to provide missing pieces of user-specific data. An identification server maintains user data that includes globally unique user identifiers. The user data can be utilized to match user actions across different online providers and to perform analyses based on matched actions.