Visited Websites May Reveal Users' Demographic Information and Personality

Cheng-You Lien, Guo-Jhen Bai, Hung-Hsuan Chen
Computer Science and Information Engineering
National Central University
littlelien.peanut@gmail.com,ivy2350442@gmail.com,hhchen@g.ncu.edu.tw

ABSTRACT

This study shows that simple supervised learning algorithms can easily predict a user's personality and demographic information based on the features derived from the users' browsing logs, even when the logs are not recorded with the finest granularity (i.e., each visited URL of a user). This is different from the analytical formula of Cambridge Analytica (CA), which reported that it needs to know each user's detailed liked objects (e.g., articles, pages, etc.) on Facebook with a fine granularity (i.e., CA needs to know the liked articles, not only the types of the articles) to predict user information. However, we employed only the visited website *categories* to predict a user's gender, age, relationship status, and big six personality scores, which is an authoritative index to represent an individual's personality in six dimensions. We also show that applying simple clustering as a preprocessing step enhances the predictive power. As a result, the data collectors, even when storing only a coarse granularity of the visited URLs of the users, may leverage such information to identify a user's preferences/tastes and her/his private information without notifying users.

ACM Reference Format:
Cheng-You Lien, Guo-Jhen Bai, Hung-Hsuan Chen. 2019. Visited Websites May Reveal Users' Demographic Information and Personality. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI '19), October 14–17, 2019, Thessaloniki, Greece.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3350546.3352525

1 INTRODUCTION

The collection of users' personally identifiable information (PII) may be at risk of breaching the EU's General Data Protection Regulation (GDPR) if such information is not properly protected. Meanwhile, Internet users have started to realize that disclosing personal information online could be dangerous; therefore, many users are reluctant to fill in online forms that request their personal information, such as gender, birthdate, education, relationship status, etc. Nevertheless, although many companies are eager to collect users' information, directly obtaining real information from users has become less straightforward.

However, a user's browsing history is typically recorded in digital format, which makes it easy for the data collector to query and make an observation. Unfortunately, it is possible to obtain a user's identity based purely on the URLs. For example, if a data collector determined that someone frequently visits the URL page https://www.linkedin.com/in/williamhgates/edit/topcard/, there is a good chance that this is the user with ID "williamhgates". One can further identify that this user is "Bill Gates" by visiting williamhgates's profile page. Additionally, a website may accidentally add personal information through query parameters in the URL. Therefore, some studies suggest replacing each URL by a pseudonym or even by a many-to-one mapping, i.e., assign one pseudonym to multiple objects, which is believed to be a safer scheme [5].

In this paper, we show that a user's personality and demographic information can be predicted by simple supervised learers when using the features derived from a user's browsing logs, even when these logs are recorded with a coarse granularity (i.e., containing only the *website type* but not the URLs). The prediction can be more accurate, when applying clustering as a preprocessing step. While this could be considered good news for data collectors, it is likely bad news for users because data collectors may characterize a user's personality and demographic information without notifying users.

This study is relevant, but different, from the approach of Cambridge Analytica (CA) (which is better known as the Facebook - Cambridge Analytica data scandal[1]) to identify a user's information. Specifically, CA's approach requires obtaining a detailed list of each user's liked objects (e.g., posts, pages, albums, etc.) on Facebook [6, 12], whereas we show that, even when the objects are recorded in a very coarse manner such that different articles may share the same identifier, one can still predict a user's personality and demographic information based on the logs with such a coarse granularity. In our experiments, we group more than 4 million distinct URLs into only 88 classes, which means that after the conversion, each identifier represents more than 40 thousand URLs on average. However, we can still identify users' personality and demographic information with such a coarse granularity of data.

2 RELATED WORK

A user's browsing history is commonly used to assist personalized recommendation and advertisement. A typical example is Google AdSense [10], which retrieves the content of a webpage to place the relevant advertisement. Additionally, Google collects personal information to customize displayed advertisements.[2] Such applications are relevant to our study because both utilize a user's browsing log

WI '19, October 14–17, 2019, Thessaloniki, Greece
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6934-3/19/10…$15.00
https://doi.org/10.1145/3350546.3352525

[1] https://en.wikipedia.org/wiki/Facebook-Cambridge_Analytica_data_scandal
[2] https://support.google.com/ads/answer/2662922

to make predictions. However, Google attempts to find the advertisement that a user is interested in, whereas we aim to predict a user's demographic information and personality test scores.

Previous studies also show that a user's visiting logs can be used to predict or identify a user's future browsing trends [1, 8, 9]. This is similar to our study because both utilize browsing logs to make predictions. However, the predicted targets are different.

A user's demographic information and personality can be predicted based on the articles he/she "liked" on Facebook [6, 12]. These studies are highly relevant to our work. However, we show that, based solely on the *category IDs* of articles a user browsed (without knowing which articles the users had read), we can still predict the personality and demographic information.

In addition to analyzing a user's interacted articles to predict private information [6, 12], involuntary information leakage may also come from the annotations from other people. Some social network platforms allow users to annotate their friends by natural language description, which may contain these friends' personal information. As a result, it is possible to infer personal information without notifying the target users [4, 7]. Although these studies are highly relevant to users' privacy leakage, they mostly utilize the online information revealed by the third party. Our predictions are based solely on the target user's browsing information.

3 DATASET

We collected 672 users' browsing logs based on a self-developed Chrome plug-in. The total number of the browsed pages among these users is 12,837,216. Table 1 shows the statistical summary of each user's number of visited pages.

Table 1: The summary of a user's number of visited pages

min	Q1	Q2	mean	Q3	max
44	4,239	13,335	19,103	26,698	130,992

In addition, we asked the users to fill out a questionnaire regarding their personal information and a big six personality test. Eventually, 513 out of the 672 users finished the personality test, and 508 out of the 513 users revealed their demographic information. We use these users' browsing behavior, demographic information, and personality test scores as our experimental dataset. The demographic information includes gender (male, female, others), age (0 - 20, 21 - 30, 31 - 40, 40+), and relationship status (single, couple, married, others). The big six personality test includes 6 dimensions: Honesty-Humility, Neuroticism, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience. The big six personality test is an extension of the traditional big five personality test, which has been recognized in modern psychology as the basic structure of all personality traits [11]. Figure 1 shows the radar chart of 3 selected users in our dataset. Each user has a unique distribution of the scores across the six dimensions.

4 PERSONALITY AND DEMOGRAPHIC INFORMATION PREDICTION

4.1 Methodology

Instead of applying the supervised learning approaches directly to predict the target variables, we propose to cluster the users first and then apply the supervised learning algorithms to each of the clusters. As we will show later in Section 5, we found that such an approach better predicts users' demographic information and personality test scores compared to applying various supervised learners directly to the entire dataset.

Let $\boldsymbol{X}_i$ denote a user i's features ($i = 1, 2, \ldots, n$); we first cluster the users into k clusters $S_1, S_2, \ldots, S_k$. We define $\boldsymbol{\mu}_j$ as the average of the user features in the cluster j; then, the objective function of a clustering algorithm can be defined by Equation 1, which attempts to minimize the sum-of-square errors from each user's feature vector to the centroid of the cluster to which this user belongs.

$$\mathcal{L} = \sum_{j=1}^{k} \sum_{\boldsymbol{X}_i \in S_j} \left\| \boldsymbol{X}_i - \boldsymbol{\mu}_j \right\|^2 . \tag{1}$$

We applied the k-means algorithm for clustering. To decide k, the number of clusters, we tested different k's and applied the Silhouette score [2] to evaluate the effectiveness. The Silhouette score $S(i)$ of the ith user can be computed by Equation 2.

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}, \tag{2}$$

where $a(i)$ is the average distance between user i and the users of the same cluster, and $b(i)$ is the average distance between user i and the users of different clusters. A positive Silhouette score means that the target user is closer to the users of the same cluster.

We applied various supervised learning methods for each cluster. For the personality test score prediction, we applied various regressors, including Lasso regression, ridge regression, elastic net regression, and support vector regression (SVR). For the demographic information prediction, we employed the following classifiers: k-nearest neighbors (KNN), logistic regression, random forests, and support vector machines (SVM).

4.2 Selected Features

We generated the features based on users' browsing logs, in which the URLs are likely the most representative information. However, we found that the distribution of users' visited URLs are highly skewed: the popular pages are visited by almost everyone (which likely makes such information a less discriminative feature), and the uncommon pages are browsed by very few individuals (which likely makes such information tend to overfit the targets). The most popular webpage, facebook.com, contributes 27.6% of the visits.

We preprocessed the URLs by categorizing the URLs based on a web classification service,[3] which converts a given URL into a corresponding web category. For example, the service converts Google "https://www.google.com" into "Search Engines and Portals". After converting all the available URLs, we obtained 88 categories, in which the top 5 popular categories and their corresponding click ratios are "Social Networking services" (29%), "Search Engines and Portals" (15%), "Email" (8%), "Media" (7%), and "Shopping" (7%).

Additionally, we found that the top 4 visited URLs belongs to Facebook, Google Search, Gmail, and YouTube, which contribute 27.6%, 11.8%, 6.6%, and 6.4% of the available clicks, respectively, which are in turn comparable to the number of clicks received by a category. Therefore, visits to these 4 URLs are treated as 4 categories.

[3] http://www.fortiguard.com/webfilter

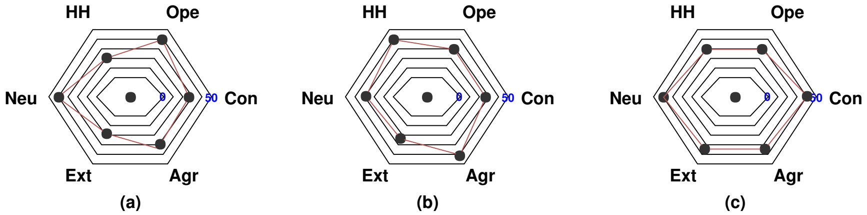

Figure 1: The radar plot of the big six personality scores of three selected users.

Eventually, we obtained 92 web categories. After the preprocessing, we generate two types of features, as described below.

The first type of features is based on an individual's overall browsing ratio on various types of pages. We define c_{ij} a user i's browsing ratio on category j by Equation 3.

$$c_{ij} = \frac{\text{user } i\text{'s number of visited pages belonging to category } j}{\text{user } i\text{'s total visited page count}}. \tag{3}$$

We believe such information is helpful in determining a user's overall interest, which may indirectly reflect a user's demographic information and personality. For example, a game-related webpage may attract more male readers who are young, whereas a bank and finance-related page may appeal to office workers.

The second type of feature is based on a user's regular browsing periods on a regular day. We believe that users with similar browsing time slots may have something in common. For example, college students tend to stay up late, so compared to other groups, college students may have active browsing behavior at midnight. We divided the time slots based on hours, so each day consists of 24 slots. We define a user i's browsing ratio on the slot j by Equation 4 (j representing the hour index of a day).

$$p_{ij} = \frac{\text{number of days having browsing behaviors in period } j}{\text{number of days having browsing behaviors}}. \tag{4}$$

4.3 Selected Targets

For the big six personality test scores, there are six scores to represent a user's six personality dimensions, which include Honesty-Humility (HH), Neuroticism (Neu), Extraversion (Ext), Agreeableness (Agr), Conscientiousness (Con), and Openness to Experience (Ope). The score of each dimension ranges between 0 and 50. Therefore, we target predicting a user's scores on these six dimensions, which can be modeled as a regression problem.

For the demographic information, we target predicting a user's gender (male, female, or other), age (16-20, 21-25, 26-30, 31-35, 36-40, 41-45, or 46 and above), and relationship status (single, in a relationship, married, others, or keep private).

5 EXPERIMENTS

5.1 Personality Prediction

For personality prediction, we predict the personality test scores in the six dimensions. Therefore, we use the root-mean-squared-error (RMSE) as the evaluation metric.

We applied four supervised regressors as the baselines, including least absolute shrinkage and selection operator (Lasso), ridge regression, elastic net model, and support vector regression. We compared the RMSE scores for all these models with and without clustering as the preprocessing step.

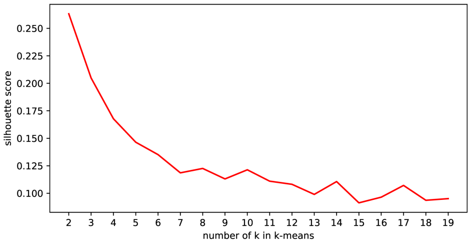

Figure 2: Cluster number vs Silhouette score.

As discussed in the explanation in Section 5.2, we fine-tuned the important hyperparameters for all the baseline models, so we believe all these compared supervised regressors represent most of their predictive power. Specifically, for the Lasso regression, ridge regression, and SVR, we tested different regularization weights on the validation set and recorded the one that yields the highest $MicroF1$ score. For the elastic net regression, we tried different regularization weights for the L1-norm and the L2-norm.

Table 2 shows the RMSE scores of all the compared supervised regressors with and without clustering as the preprocessing step. As shown, all the supervised regressors improve (i.e., smaller RMSE) when we cluster the users first. However, there seems to be no obvious winner among the four supervised learners.

We found that if we do not cluster the users first, the Lasso regressor, ridge regressor, and elastic net regressor perform better when the regularization weight is very large. This suggests that these models tend to minimize the norms of the parameters to learn instead of minimizing the training error. As a result, these models' predictive power may be close to the naïve average model, which always returns the average of the target variable in the training data as the prediction. However, if we cluster the users first, the best performing models may end up having smaller regularization weights. As a result, the models likely obtain more clues between the features and the target variables in the training data.

5.2 Demographic Information Prediction

As we discussed in Section 4.3, we model the task of demographic information prediction as a classification problem. For the binary classification problem, the $F1$ score is widely used as an evaluation metric because the $F1$ score considers the harmonic mean of the precision and the recall scores. However, the regular $F1$ score formula may not work for the multiclass classification problem. As a result, we instead used the $MicroF1$ score as the evaluation metric [3]. The $MicroF1$ score, as shown by Equation 5, is defined

Table 2: Comparing the pure supervised regressors with our proposed method (clustering preprocessing + supervised classifier), based on the RMSE score on the test dataset.

Method	Supervised regressor						Clustering + supervised regressor					
Prediction target	HH	Neu	Ext	Agr	Con	Ope	HH	Neu	Ext	Agr	Con	Ope
Lasso	5.832	5.87	5.881	5.71	5.406	5.607	**5.411**	**5.469**	**5.435**	**5.435**	**5.022**	**5.131**
Ridge	5.845	5.981	5.891	5.795	5.43	5.646	**5.43**	**5.404**	**5.38**	**5.325**	**5.027**	**5.052**
Elastic net	5.813	5.769	5.743	5.622	5.366	5.44	**5.417**	**5.383**	**5.422**	**5.317**	**5.022**	**5.095**
SVR	5.789	5.78	5.746	5.643	5.232	5.38	**5.432**	**5.623**	**5.402**	**5.328**	**5.048**	**5.165**

Table 3: Comparing the pure supervised classifiers with our proposed method (clustering preprocessing + supervised classifier), based on the *MicroF*1 score on the test dataset.

Method	Supervised classifier			Clustering + supervised classifier		
Prediction target	Age	Gender	Relationship	Age	Gender	Relationship
Baseline	0.388	0.545	0.474	**0.411**	**0.598**	**0.476**
KNN	0.427	0.594	0.478	**0.435**	**0.618**	**0.482**
Random forest	**0.453**	**0.697**	0.488	0.419	0.687	**0.512**
Logistic regression	0.427	**0.697**	0.476	**0.457**	0.675	**0.498**
SVM	0.388	0.591	0.474	**0.411**	**0.642**	**0.512**

as the harmonic mean of the micro-average of the precision score *MicroP* and the micro-average of the recall score *MicroR*, which are defined by Equation 6 and Equation 7, respectively.

$$MicroF1 = \frac{2MicroP \cdot MicroR}{MicroP + MicroR}. \tag{5}$$

$$MicroP = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} (TP_i + FP_i)}, \tag{6}$$

where C is the number of classes, TP_i is the number of true positives when regarding the class i as the positive class and the others as negative, and FP_i is the number of false positives when regarding the class i as the positive class and the others as negative.

$$MicroR = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} (TP_i + FN_i)}, \tag{7}$$

where FN_i is the number of false negatives when regarding the class i as the positive class and the others as the negative class.

We selected the *MicroF*1 score instead of the *MacroF*1 [3] score because the *MacroF*1 score does not consider the issue of imbalanced dataset. As a result, the *MacroF*1 score is very sensitive to the precision and recall of the classes with a small number of cases.

We selected the k-nearest neighbors algorithm, logistic regression, random forest, and support vector machines as the baseline classifiers for comparison.

For each method, we carefully tuned the important hyperparameters. As a result, each baseline method shows most of its predictive power. Specifically, for KNN, we tested different k's on the validation set and selected the one that yielded the best *MicroF*1 score. We also tested different regularization weights for the logistic regression classifier and the SVM classifier and different tree depths for the random forest classifier.

Our proposed method – clustering before classification – requires deciding the number of clusters. We leverage the Silhouette score to decide the number of clusters. As shown in Figure 2, when the number of clusters is less than 6, the Silhouette score is apparently larger than the rest. Therefore, we tested the cluster numbers from 2 to 6 on the validation set for all the experiments.

Table 3 shows the *MicroF*1 scores of the supervised classifiers with and without clustering as the preprocessing step. Applying clustering yields better results in most cases. However, there seems to be no obvious winner among the various supervised classifiers.

6 DISCUSSION

This paper shows that users' personality traits and demographic information can be predicted based on the browsing logs, even when the URLs in the logs are preprocessed by a many-to-one pseudonym, which is believed to be a safer scheme to protect the identity of the objects. With standard supervised classifiers or regressors, a user's gender, age, relationship status, and personality test scores can be predicted. We also show that by clustering the users into groups and applying supervised learning algorithms to each group independently, the predictions can be more accurate. Therefore, we suggest that even when safeguarding the release of anonymized logs with a many-to-one pseudonym on the visited URLs, it is still necessary to inject random noises to preserve privacy.

For future work, we are interested in investigating the relationship between private information and unconscious behaviors, such as the frequency of mouse clicks and typing speed. A user may be able to intentionally browse different types of webpages to disguise his/her interest and personality in a short period, but unconscious behaviors are less likely to change. If these unconscious behaviors can indeed be used to predict users' private information, this suggests that users' privacy may be compromised by merely recording users' clicking and typing frequency.

ACKNOWLEDGMENTS

We acknowledge partial support by the Ministry of Science and Technology under Grant No.: MOST 107-2221-E-008-077-MY3. We are grateful to the National Center for High-performance Computing for computer time and facilities. We appreciate Dr. Tsung-Ren Huang and his team for developing the first version of the Google Chrome plug-in for this project.

REFERENCES

[1] Guo-Jhen Bai Bai, Cheng-You Lien Lien, and Hung-Hsuan Chen. 2019. Co-learning Multiple Browsing Tendencies of a User by Matrix Factorization-based Multitask Learning. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE.

[2] Renato Cordeiro de Amorim and Christian Hennig. 2015. Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences* 324 (2015), 126–145.

[3] Nadia Ghamrawi and Andrew McCallum. 2005. Collective multi-label classification. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. ACM, 195–200.

[4] Ralph Gross and Alessandro Acquisti. 2005. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society*. ACM, 71–80.

[5] Saroj Kaushik, Shivendra Tiwari, and Priti Goplani. 2011. Reducing dependency on middleware for pull based active services in LBS systems. In *International Conference on Wireless Communications and Applications*. Springer, 90–106.

[6] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* (2013), 201218772.

[7] Ieng-Fat Lam, Kuan-Ta Chen, and Ling-Jyh Chen. 2008. Involuntary information leakage in social network services. In *International Workshop on Security*. Springer, 167–183.

[8] Chen-You Lien, Guo-Jhen Bai, Ting-Rui Chen, and Hung-Hsuan Chen. 2018. Predicting User's Online Shopping Tendency During Shopping Holidays. In *The 2018 Conference on Technologies and Applications of Artificial Intelligence*.

[9] Caroline Lo, Dan Frankowski, and Jure Leskovec. 2016. Understanding behaviors that lead to purchasing: A case study of pinterest. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 531–540.

[10] Dwight Allen Merriman and Kevin Joseph O'connor. 1999. Method of delivery, targeting, and measuring advertising over networks. (Sept. 7 1999). US Patent 5,948,061.

[11] Brian P O'Connor. 2002. A quantitative review of the comprehensiveness of the five-factor model in relation to popular personality inventories. *Assessment* 9, 2 (2002), 188–203.

[12] Wu Youyou, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences* 112, 4 (2015), 1036–1040.