Customer Gender Prediction Based on E-Commerce Data

Duc Duong, Hanh Tan
Faculty of Information Technology
Posts and Telecommunications Institute of Technology
Hanoi, Vietnam
{ducdt, tanhanh}@ptit.edu.vn

Son Pham
Faculty of Information Technology
University of Engineering and Technology
Hanoi, Vietnam
sonpb@vnu.edu.vn

***Abstract*— Demographic attributes of customers such as gender, age, etc. provide important information for e-commerce service providers in marketing and personalization of web applications. However, online customers often do not provide this kind of information due to privacy issues and other security-related reasons. In this paper, we proposed a method for predicting the gender of customers based on their catalog viewing data on e-commerce systems, such as the date and time of access, list of categories and products viewed, etc. We employ a machine learning approach and investigate a number of features derived from catalog viewing information to predict the gender of viewers. Experiments were conducted on datasets provided by the PAKDD'15 Data Mining Competition and achieved the good result. The results 81.2% on balanced accuracy and 81.4% on macro F1 score showed that basic features such as viewing time, products/categories features used together with more advanced features such as products/categories sequence and transfer features effectively facilitate gender prediction of customers.**

Keywords*— machine learning, big data, demographic prediction

I. INTRODUCTION

Today, many web applications such as e-commerce systems, search engines, online marketing systems, employ personalization features to increase the user experience. With a good personalized service, the information displayed is customized for each user individually rather than remaining the same for all users. For example, e-commerce systems can display promotions or recommend products which are relevant to the individual visitor rather than random promotions or products.

Personalization is mainly based on two types of data: historical data (e.g. previous items viewed or purchased) and demographic attributes of users (e.g. gender, age, education etc.). Historical data can be obtained only if the user has used the system before and has logged into the system. Therefore, historical data-based methods are unusable for guest or new users. Demographic-based methods are useful even when the user has never used the system before. However, this information is not easy to obtain, because Internet users are often not willing to provide their private information. For that reason, in many cases, the only way to obtain the demographic attributes of users is to predict. The task of author profiling for anonymous texts are studied for decades ([1], [2], [4], [6], [8], [10], [15], [17]), but not all users write something on the system. Another way to predict demographic information of users is based on their behaviors on the systems, for example browsing activities ([7], [14]), website traffic ([3]), or catalog viewing data. The main advantage of this approach is that data is available in most cases, because users must do something on the system such as access pages, click items, or browse the catalog.

In this research, we address the problem of predicting demographic information of users based on their catalog viewing data such as viewing time/duration, viewed categories/products, etc. Datasets for our experiment were provided by FPT Corporation in the PAKDD'15[1] Data Mining Competition. Beside basic features such as time and duration of viewing, individual products/categories viewed in the session, we investigate features which contain information about relation of products/categories viewed in the session, such as products/categories sequence and transfer, etc. (we call them "advanced features"). With multi-level hierarchical structure of categories/products, we employed a tree-based feature representation which provides a better view for feature extraction than the list-based style. Popular learning methods such as Random Forest, Support Vector Machine (SVM), and Bayesian Network (BayesNet) were used for experiments and we also employed supporting techniques for dealing with the class-imbalance problem such as resampling, cost-sensitive learning to improve overall prediction accuracy. Results are promising although more types of feature and techinique need to be investigated to improve the performance.

The organization of the paper is as follows. In section II, we present related work on user demographic prediction. Section III describes methods and the system. Section IV presents results and discussion. In section V, we draw a conclusion and future work.

II. RELATED WORK

Demographic prediction has been studied for a long time. At the early stage, most of researches on this field focused on

[1] PAKDD'15: Pacific-Asia Conference on Knowledge Discovery and Data Mining 2015

978-1-4673-8929-7/16/$31.00 ©2016 IEEE

authorship studies, which are tasks of determining or predicting author characteristics by analyzing texts created by him/her. Methods which researchers used in these studies are mostly based on analysis of writing style using various types of features, such as lexical, syntactic, or content-based features. The first study in this field dates back to 19th century when Mendenhall (1887) investigated Shakespeare's plays. But the work which was considered the most thorough study in this field was conducted by Mosteller and Wallace (1964) when they analyzed the authorship of FederalList Papers based on Bayesian statistical analysis of frequencies of function words. The previous authorship studies often focused on literature texts, such as novel or article. Recently, due to the growth of Internet and online communication channels, the focus has been moved to computer mediated communication contents, such as email, blogs, comments, etc. Reference [4] used 221 features to determine the authorship of emails. Reference [1] investigated the differences in writing style of male and female in 604 documents from British National Corpus. Reference [2] explored the use of stylometric and content-based features to predict gender and age of bloggers on datasets with over 71.000 blog posts from blogger.com. This model achieved results 80% for gender prediction and 76% for age prediction. Reference [8] proposed a method to calculate a "write print" based on frequent patterns extracted from emails to predict to profile of the author. Reference [15] conducted a research to predict gender and age of twitter messages and forum posts using regression methods with accuracy around 80%.

Beside the methods based on the analysis of textual data, recently, researchers investigated the use of user behaviors on web applications to predict their demographic information. Reference [7] proposed a method to solve the problem of predicting Internet users' gender and age based on their browsing behaviors. They used webpage view information as input variables to propagate demographic information of users. The SVM method was employed on features set consisting of content-based (words from the Webpages) and category-based (hierarchy of web concepts) features and achieved the results of 79.7% on gender and 60.3% on age. Reference [9] also investigated machine learning approaches to predict the demographic attributes of websites using information from the content and hyperlinked structure. The research of [5] aimed at inferring users' demographics based on their daily mobile communication patterns. Their study was conducted on a real-world large mobile network of more than 7.000.000 users and over 1.000.000.000 communication records. They used the features set including individual features, friend features, and circle features and achieved the best results of 80% for gender and 70% for age. Reference [19] proposed a prediction framework for predicting users' demographics based on the analysis of their behaviors and environments. They also developed a new method namely Multi-Level Classification Model to solve the imbalanced class problem. Reference [14] also addressed the problem of predicting users' gender based on browsing history. They employed classification-based methods and used features derived from browsing log data. They added more types of features than the previous works, such as topic-based features, time features, sequential features, and improved results significantly.

In this paper, we investiged a method for predicting user demographics based on their catalog browsing behaviors on e-commerce systems. As far as we have known, there is no thorough research work on this problem.

III. Approach

A. System overview

In this work, we developed a system which can take data from product viewing logs for users with known gender, extract features and class labels to create a training dataset. A model is built from the training dataset using a classification-based method and then can be used to predict the gender of unknown users based on their product viewing activities.

The training data file contains records which correspond to product viewing logs. A single log contains information about products viewing data of a user, such as session start time, session end time, list of products and categories IDs. The class labels for each training sample are male and female. Therefore, the task is a binary classification problem with two labels correspondingly.

In the next sections, we describe features and techniques which were used for prediction in detail.

B. Features

The feature set we used in this work can be divided into two types, which we call basic and advanced features.

1) Basic features

Basic features include temporal and individual products/categories features. Temporal features are features related to timestamp and frequency of viewing activities. Time in day, day of week, holidays, viewing duration, number of products viewed in one session, etc. are the factors that can be used to predict the gender of a customer. We used totally 98 binary and 3 numeric features of this kind as shown in Table I.

TABLE I. TEMPORAL FEATURES

Features	Description
Day	Day in month (31 features)
Month	Month in year (12 features)
DayOfWeek	Day in week (7 features)
StartTime	Exact hour (24 features)
EndTime	Exact hour (24 features)
Duration	Session length (1 features)
NumberOfProducts	Number of products (1 features)
AverageTimePerProduct	Average viewing time of a product (1 features)

Individual products/categories features consist of all categories and products in the system. Because provided datasets contain all the categories and products IDs, we just extracted them and used as features. For each category or product, we count the number of times the user has browsed it and used that number as the feature value. As each complete product ID can be decomposed into four different IDs, from the most general categories (start with "A") to the subcategories (start with "B" and "C") and individual product (start with "D") respectively, we have 4 types of features this kind, with 2.057 features in total as shown in Table II. We note that due to

the large number of individual product IDs, we only choose the IDs which appear at least 3 times.

TABLE II. INDIVIDUAL PRODUCTS/CATEGORIES FEATURES

Features	Description
The most general categories	IDs start with A (11 features)
Subcategory level 1	IDs start with B (60 features)
Subcategory level 2	IDs start with C (186 features)
Product	IDs start with D (1.800 features)

1) Advanced features.

Beside individual categories/products features, we hypothesize that the relation between categories/products viewed during a single session also reflects gender of the viewer. The categories/products viewed in a session are presented in the list-based style as the following:

A00002/B00003/C00006/D19760/;
A00002/B00001/C00010/D18416;
A00002/B00001/C00004/D19764/;A00002/B00003/C00008/D19761/;
A00002/B00003/C00008/D08538/

Because the list-based presentation may cause the difficulties for extracting all the relation information between individual categories/products, we proposed a tree-based presentation, in which the most general category is the root of the tree, the products are at the leaves of tree, and the subcategories are placed at intermediate levels. For example, the above list-based presentation of categories/products can be converted to a tree-based presentation as in the Fig. 1.

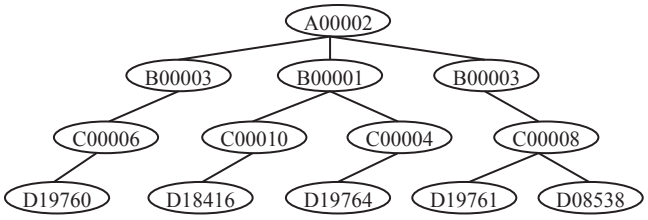

Fig. 1. Tree-based presentation of categories/products viewed in a session.

From that tree, we can easily obtain the list of categories/products by travelling the tree in depth and from the leftmost. Moreover, from the tree view, we can extract the relation information between categories/products by exploring properties of the tree such as nodes, levels, paths, siblings, etc. For our problem, we can use the following properties of tree as features:

- Number of nodes at each level
- Sequences of nodes at the same level: From the sequence of nodes at each level, we extract all k-grams subsequence of it and choose the most frequent k-grams
- Node transfer pairs at different levels: These features reflect the browsing habit of users when moving to another category at the different level.

For example, in the above tree, these properties can be extracted as the following:

- Number of nodes at each level are {1, 3, 4, 5}
- Sequences of nodes with the same level are {B00001, B00003, B00001}, {B00001, B00003}, {C00006, C00010}, {D19760, D18416, D19764}, etc.
- Node transfer pairs are {D19760, B00001}, {D18416, C00004}, etc.

Because of large number of categories and products, the total possible sequences and transfer pairs may be very huge. Therefore, we only choose the sequences and pairs which appear at least 3 times in the training dataset. By that way, we have number of advanced features as shown in Table III.

TABLE III. ADVANCED FEATURES

Features	Description
Number of nodes at each level	4 features
Most frequent sequences of categories/products	1.100 features
Most frequent node transfer pairs at different levels	300 features

C. Classification techniques

We used three machine learning algorithms Random Forest, SVM, and BayesNet to learn the model. However, because the training data is imbalanced (around 80% of females and 20% of males), we employed some supporting techniques such as Cost-Sensitive Learning, Resampling, Class balancing to improve the accuracy.

Resampling methods are commonly used for dealing with class-imbalance problem. The basic idea is to add or remove some instances to make the dataset become more balanced. Therefore, there are two methods of resampling that are under-sampling (reduce the number of large class) and over-sampling (replication of small class instances). According to [11], the main drawback of under-sampling is that this method can discard the potential data which can be important for the task, while over-sampling may lead to additional computation cost and over-fitting problem in case of random replication. Another approach to resampling is class balancing technique, which reweights the instances in each class to obtain a same total class weight, instead of duplicating or eliminating instances. In our experiments, we used the class balancing method in combination with the Cost-Sensitive Learning algorithm.

While resampling is the data-level method, cost-sensitive learning is an algorithm-level method to solve the problem of class-imbalance classification [11]. According to [12], cost-sensitive learning is a method that takes the misclassification cost into the consideration, meaning it treats the different misclassifications differently.

Lastly, because the total number of features is still quite large (3.500 features), we apply a feature selection technique to reduce the complexity and eliminate the features which are not discriminating. In our work, we used the Information Gain to select 2.500 features which have highest mutual information.

IV. EXPERIMENTS

A. Data

We used datasets from the PAKDD'15 Data Mining Competition which were provided by FPT Corporation. The data was divided into training and test sets. Each of set contains 15,000 records which correspond to the product viewing logs.

A single log in the training data file is composed of four types of information:

- Session ID
- Start time (including date)
- End time (including date)
- List of product IDs

The list of product IDs contains the IDs of the products which the user has viewed during the session. Because the products may belong to different categories, the information about categories is also included in IDs. An example of a single log is as follow:

u10008, 2014-11-17 19:20:06, 2014-11-17 19:21:54, A00001/B00001/C00001/D00001/; A00001/B00002/C00002/D00002

B. Evaluation metrics

As mentioned earlier, due to the class-imbalance problem, the balanced accuracy measure (BAC) was used to evaluate the model. Balanced accuracy is defined as an average accuracy obtained on either class and can avoid inflated performance estimates on imbalanced datasets.

$$balanced\ \ accuracy = \frac{0.5 * tp}{tp + fn} + \frac{0.5 * tn}{tn + fp}$$

Where *tp* is true positives, *tn* is true negatives, *fp* is false positives, and *fn* is false negatives.

This measure was also used to evaluate the results in the PAKDD'15 Data Mining Competition.

In this work, we report this score together with macro F1 score to facilitate the comparison with previous works.

C. Results and Discussion

In order to evaluate the performance of basic and advanced features, we conducted experiments on different sets of features, including basic features only and combination of both types of features. Each set of features was tested using the popular machine learning methods, namely Random Forest, SVM, and BayesNet. The training data and testing datasets are provided separately (each dataset has 15.000 samples). Therefore, our model was created based on the training dataset and tested on a different dataset. In addition, to verify the effects of supporting techniques such as Cost-Sensitive Learning, Resampling, Class balancing, we experimented on various combination and found that using cost-sensitive learning solely with cost matrix 1:4 achieved best Marcro F1 score (81.4%) but using in combination with class balancer filter with cost matrix 1:3 gave best BAC score (81.2%). Table IV-V shows the results of our experiments.

TABLE IV. RESULTS OF EXPERIMENTS ON COST-SENSITIVE LEARNING WITH CLASS BALANCING

	Basic features only		**Basic + Advanced features**	
	BAC	***Macro F1***	***BAC***	***Macro F1***
Random Forest	77.5	75.8	**81.2**	**78.8**
SVM	76.8	74.6	79.8	77.0
BayesNet	76.2	74.8	78.8	76.2

TABLE V. RESULTS OF EXPERIMENTS WITH COST-SENSITIVE LEARNING ONLY

	Basic features only		**Basic + Advanced features**	
	BAC	***Macro F1***	***BAC***	***Macro F1***
Random Forest	76.6	77.4	**80.8**	**81.4**
SVM	76.0	76.2	79.3	78.8
BayesNet	75.2	75.8	78.2	78.6

As the results shown in Table IV, when using cost-sensititve learning with class balancing, Random Forest achieved the best results while BayesNet gave the lowest performance on both BAC and Macro F1 score, in which the best BAC score (81.2%) is better than Macro F1 score (78.8%). But the Table V shows that when using cost-sensitive learning only, the best Macro F1 score increased to 81.4%, while BAC score reduced to 80.8%.

The advanced features when combining with basic features also remarkablely improve prediction result compared with using basic features only. However, in provided datasets, there are many sessions in which users only view one product and the advanced features don't have any affect on these cases. In fact, users often view more than one products when surfing on e-commerce systems. Therefore, we believe that the improvement will increase more when applying our method on more realistic datasets.

For the number of features, we selected 2.500 because when we conducted the experiments with different number of features ranging from 1.000 to 3.500, we found that the prediction result increases and reaches the top at 2.500 features. Fig. 2 shows the BAC scores of the method when using different number of features.

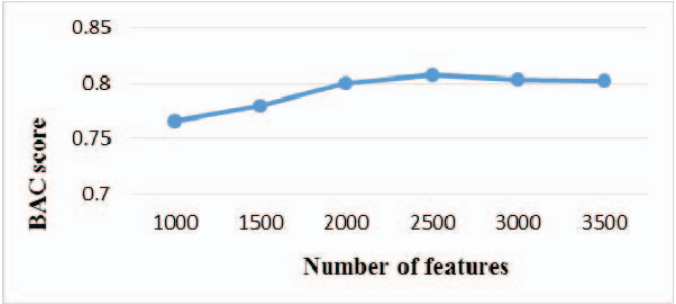

Fig. 2. BAC scores of different number of features

D. Comparison with the earlier works.

The baseline results of demographic prediction tasks based on text analysis is 80% for gender (accuracy). Therefore, the BAC and Macro F1 score of gender prediction in our work can be considered promising. With more similar works conducted by [7] and [14], which predicted users' gender based on their web browsing data, the F1 score of our work is also good, although the browsing activities generate more meaningful data for training than product viewing activities. Moreover, web pages contain words, therefore, they can use more types of features such as content words or page topics, etc.

In comparison with other solutions from the teams participating in PAKDD'15 Data Mining Competition, the result of this work is in the top 10 of leader scores. The performance of the best team is 87.9%, and the top 10 positions achieved the results from 81%. However, for the purpose of getting the highest possible score in the competition, most of their solutions used the datasets-specific features such as product ID prefix, session alignment between sessions in training and test sets, etc. Therefore, those solutions may not get that good results when apply to other datasets. In this work, we investigate a general solution which can be applied on any dataset, therefore, we have not tried to use these types of features.

V. Conclusion

In this study, we investigated a method for predicting the gender of customer based on product viewing data on e-commerce systems. We proposed an approach which used basic features such as viewing time and duration, individual categories/products in combination with advanced features such as categories/products sequences and transfer pairs, which we extract from a tree-based presentation of categories /products list. This feature design is work best on the Random Forest algorithm with supporting techniques such as Cost-Sensitive Learning, Resampling, Class balancing to deal with class-imbalance problem. In addition, the advantage of method is that it can be easily applied to other datasets because it uses no dataset-specific features.

In the future, this work can be investigated further on the feature set. More type of features can be inferred from the tree-based presentation to exploit the relation between the products viewed in the same session. We also have plan to collect data from other sources to improve the general performance and expand to other demographic attributes such as age, location, job, and so on.

VI. References

[1] S. Argamon, M. Koppel, J. Fine, and A. Shimoni, "Gender, genre, and writing style in formal written texts," Text 23(3), August 2003.

[2] S. Argamon, M. Koppel, J. Pennebaker, and J. Schler, "Automatically profiling the author of an anonymous text," Communications of the ACM , v.52 n.2, February 2009.

[3] J. C. A. Culotta, N. R. Kumar, and J. Cutler, "Predicting the demographics of twitter users from website traffic data, " Proceedings of the 29th AAAI Conference on Artificial Intelligence, Jan 2015.

[4] O. De Vel, A. Anderson, M. Corney, and G. M. Mohay, "Mining e-mail content for author identification forensics," SIGMOD Record 30(4), pp. 55-64, 2001.

[5] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla, "Inferring user demographics and social strategies in mobile social networks." In: KDD'14. ACM. p. 15–24, 2014.

[6] D. T. Duc, P. B. Son, and T. Hanh, "Using content-based features for author profiling of Vietnamese forum posts," In: Recent Developments in Intelligent Information and Database Systems, pp. 287–296. Springer International Publishing, Berlin, 2016.

[7] J. Hu, H. J. Zeng, H. Li, C. Niu, and Z. Chen, "Demographic prediction based on user's browsing behavior," Proceedings of the 16th international conference on World Wide Web, pp. 151-160, 2007.

[8] F. Iqbal, M. Debbabi, B. C. M. Fung, and L. A. Khan, "E-mail authorship verification for forensic investigation," Proceedings of the 2010 ACM Symposium on Applied Computing, ser. SAC '10. New York, NY, USA: ACM, pp. 1591-1598, 2010.

[9] S. Kabbur, E. H. Han, and G. Karypis, "Content-based methods for predicting web-site demographic attributes," Proceedings of ICDM, pp. 863-868, 2010.

[10] M. Koppel, S. Argamon, and A. R. Shimoni, "Automatically categorizing written texts by author gender," Literary and Linguistic Computing, 17(4), pp : 401-412, 2002.

[11] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling unbalanced datasets: A review," GESTS International Transactions on Computer Science and Engineering 30 (1), pp. 25-36, 2006.

[12] C. X. Ling, and V. S. Sheng, "Cost-sensitive learning and the class imbalance problem." In: Sammut C (ed) Encyclopedia of machine learning. Springer, Berlin, 2008.

[13] M. Pennachiotti, and A. M. Popescu, "A machine learning approach to Twitter user classification". Proceedings of AAAI, 2011.

[14] T. M. Phuong, and D. V. Phuong, "Gender prediction using browsing history," Proceedings of the Fifth International Conference KSE 2013, Volume 1. pp. 271-283, 2013.

[15] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder, "How old do you think i am?; a study of language and age in twitter," Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, 2013.

[16] F. Rangel, and P. Rosso, "Use of language and author profiling: Identification of gender and age," In Natural Language Processing and Cognitive Science, p. 177, 2013.

[17] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker, "Effects of age and gender on blogging," In Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, pp. 191-197, 2006.

[18] R. E. Schapire, "The boosting approach to machine learning: An overview," Proc. MSRI Workshop Nonlinear Estimation and Classification, 2001.

[19] J. J. C. Ying, Y. J. Chang, C. M. Huang, and V. S. Tseng, " Demographic prediction based on users mobile behaviors," In Nokia Mobile Data Challenge, 2012.

[20] C. Zhang, and P. Zhang, "Predicting gender from blog posts," Technical Report. University of Massachusetts Amherst, USA, 2010.