IFERP
connecting engineers... developing research

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 4, Issue 5, May 2017

Client Gender Prevision based on the E-commerce Information

[1] Srinivasa R B.E, M.Tech (Ph.D), [2] Deepika M
[1] Associate professor, Department of computer science, RRCE, VTU, Bengaluru..
[2] Student, Department of computer science, RRCE, VTU, Bengaluru.

***Abstract*— The Demographic property of clients such as gender, age, and education gives important information for e-commerce service providers or specialists in merchandising and personalization of web applications. However, online clients often do not give this type of information due to privacy and security related reasons. In this we proposed a method for previsioning the gender of clients based o0n their catalogue viewing data on e-commerce systems, such as the date and time of access, list of categories and products viewed, etc. We use a machine learning techniques and investigate a number of characteristics derived from catalogue viewing information to prevision the gender of viewers. Experiments were carried out on the datasets. The results 81.2% on balanced accuracy shows that basic characteristics such as viewing time, products/categories characteristics used together with more advanced characteristics such as products/categories sequence and transfer characteristics effectively facilitate gender prevision of clients.**

Keywords – Big data, classification techniques, machine learning algorithms, demographic properties.

I.INTRODUCTION

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

Personalization is based on the two types of data: that is the historical data for example previous items viewed or purchased and demographic attributes of users for example gender, age, education etc. Historical data can be obtained only if the user or client has used the system before and logged into the system. Therefore, historical data-based methods are unusable for guest, client or new users. Demographic-based methods are useful even when the user or client has never used the system before. However, this information is not easy to get, because Internet users are often not willing to provide their personal data, due to this reason, in many cases, the only way to obtain the demographic attributes of users is to prevision. One way to predict demographic data of users or clients is based on their behaviours on the systems, for example surfing activities ([1], [2]), website traffic ([3]), or catalogue 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 catalogue.

In this we deal with the problem of previsioning demographic data of users or clients based on their catalogue viewing data such as viewing time/duration, viewed categories/products, etc. The basic characteristics such as time and duration of viewing, individual products/categories viewed in the session and also we enquire characteristics which contain information about relation of products/categories viewed in the session, such as products/categories sequence and transfer, etc. we call them as “advanced characteristics”. With multi-level hierarchical structure of categories/products, a tree-based characteristic representation which provides a good view for characteristic excerption than the list-based style of representation. The most popular learning methods such as Random Forest, Support Vector Machine (SVM), and Bayesian Network (Bayes Net) [4] which deal with techniques for the prevision are used for experiments on datasets and also for solving the class-imbalance problem supporting techniques such as resampling, cost-sensitive learning to improve overall prediction accuracy.

The paper is organized in the following way. The related work on the user or client demographic prevision is presented in section II. Methods and the system used for the proposed work is provided in section III. The results and discussion is presented in section IV. In section V gives conclusion and future work of this paper.

II. RELATED WORK

Demographic prevision has been studied for a long time. At the early stage, most of researches on this field focused on authorship studies, which are tasks of determining author characteristics by analysing texts created by him/her. Methods which researchers used in these studies are mostly based on analysis of writing style using various types of, such as lexical, syntactic, or content-based characteristics [5].

In paper “Gender, genre, and writing style in formal written texts[3,10],” the results presented offer convincing evidence that there are indeed different strategies employed by men and women in setting forth information and especially in encoding the relation between writer and reader in texts Ascertaining the precise communicative functions and broader social significance of these respective linguistic strategies is a difficult and ideologically-loaded problem which is beyond the scope of this paper. Nevertheless, the fact that these results extend findings substantiated independently in less formal communication contexts to large formal written texts intended for an unseen audience over a range of genres is very suggestive [7]. The extension to low-interaction linguistic modalities invites a re-examination of the mechanisms of socialization of men and women into interactional styles and related differences in the use of language and hints at the possibility that new learning and other cognitive explanations may be called for and in the paper “Predicting the demographics of twitter users from website traffic data[8, 9], ”we have shown that Twitter follower information provides a strong source of information for performing demographic inference. Furthermore, pairing web traffic demographic data with Twitter data provides a simple and effective way to train a demographic inference model without any annotation of individual profiles. We have validated the approach both in aggregate (by comparing with Quant cast data) and at the individual level (by comparing with hand labeled annotations), finding high accuracy in both cases. Somewhat surprisingly, the approach outperforms a fully supervised approach for gender classification, and is competitive for ethnicity classification.

In this paper “Inferring user demographics and social strategies in mobile social networks [10, 11],” they studied the human interactions on demographics by investigating a country-wide mobile communication network. From this, we discover a set of social strategies stemming from human communications. First, young people put more focus on enlarging their social circles; as they age, they have the tendency to maintain small but close social relationships. Second, we observe a strong homophile of human interactions on gender and age simultaneously. Third, beyond these observations, we find that the frequent cross-generation interactions are maintained to pass the torch of family, workforce, and human knowledge from generation to generation in social society [15]. Finally, we observe striking gender differences in social triadic relationships across individuals’ lifespans, which reflects dynamic gender-bias of human behaviors from young to old.

The paper “Demographic prediction based on user’s browsing behaviour [12, 13],” focuses on demographic prevision based on people’s internet browsing history. A novel solution is proposed to train a gender and age prevision based on web users’ browsing behaviors. Experimental results on a real large page click-through log indicate that our proposed algorithm can achieve 79.7% on gender and 60.3% on age in terms of Macro F1, which achieves up to 30.4% improvements on gender prevision and 50.3% on age prevision in terms of macro F1, comparing with baseline algorithms.

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 [9]. 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 [15].

The training data file contains records which corresponding 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 [16].

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

B. Characteristics

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

- Basic characteristics

Basic characteristics include temporal and individual products/categories characteristics. Temporal characteristics are characteristics 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. There are totally 98 binary and 3 numeric characteristics of this kind.

TEMPORAL CHARACTERISTICS

- Day in month (31 characteristics)
- Month in year (12 characteristics)
- Day-Of-Week Day in week (7 characteristics)
- Start Time Exact hour (24 characteristics)
- End Time Exact hour (24 characteristics)
- Duration Session length (1 characteristics)
- Number of Products (1 characteristics)
- Average Time per Product (Average viewing time of a product) (1 characteristics)

Individual products/categories characteristics 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 characteristics. For each category or product, we count the number of times the user has searched it and used that number as the characteristic value. As each complete product ID can be decomposed into four different IDs, from the most general categories (start with “P”) to the subcategories (start with “Q” and “R”) and individual product (start with “S”) respectively [17, 18], we have 4 types of characteristics this kind, with 1500 characteristics in total, due to the large number of individual product IDs, we only choose the IDs which appear at least 4 times.

INDIVIDUAL PRODUCTS/CATEGORIES CHARACTERISTICS

- General categories IDs start with P (10 characteristics)
- Subcategory level 1 IDs start with Q (60 characteristics)
- Subcategory level 2 IDs start with R (180 characteristics)
- Product IDs start with S (1,750 characteristics
- Advanced characteristics.

Beside individual categories/products characteristics, 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:

P00004/Q00001/R00008/S17750/;
P00004/Q00005/R00012/S16492/;
P00004/Q00005/R00003/S18862/;
P00004/Q00002/R00006/S18765/;
P00004/Q00002/R00006/S09345/;

Because the list-based representation may cause the difficulties for extracting all the relation information between individual categories/products, we proposed a tree-based representation, 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 list-based representation of categories/products can be converted to a tree-based representation as in the Figure 1.

From the tree, we can easily obtain the list of categories/products by analysing 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 characteristics:

- 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.

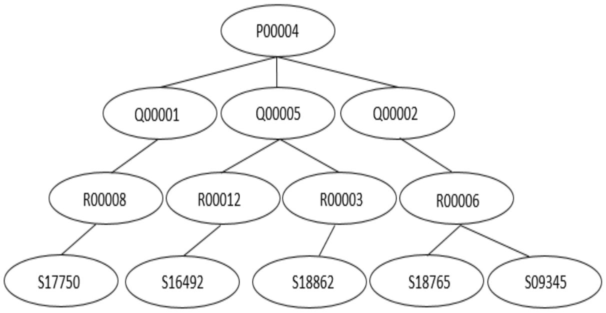


Fig. 1. Tree-based presentation of categories/products viewed in a session.

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 {Q00001, Q00005, Q00002}, {Q00001, Q00005}, {R00008,R00012}, {S17750, S16492, S18862}, etc.
- Node transfer pairs are {S17750, Q00001}, {S16492,Q00005} 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 4 times in the training dataset.

ADVANCED CHARACTERISTICS

- Number of nodes at each level has 4 characteristics.
- Most frequent sequences of categories/products has 800 characteristics.
- A most frequent node transfer pair at different levels has 200 characteristics.

C. Classification methods

We used three machine learning algorithms Random Forest, SVM (Support Vector Machine) and Bayes-Net [4] to learn the model and we used some supporting techniques such as Cost-Sensitive Learning [17] and Resampling [18] 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). 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.

Resampling is the data-level method, cost-sensitive learning is an algorithm-level method to solve the problem of class-imbalance classification [18]. According to [17], 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 characteristics is still quite large (2,600 characteristics), we apply a characteristic selection technique to reduce the complexity and eliminate the characteristics which are not discriminating. In our work, we used the Information Gain to select 1,500 characteristics which have highest mutual information.

IV. EXPERIMENTS

A. Data

The data is divided into training and test sets. Each of set contains 10,000 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 with date
- End time with 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:

u10002, 2015-12-04 11:16:30, 2015-12-04 01:05:20,
P00001/Q00002/R00003/S00004/;
P00001/Q00002/R00003/S00004/;

B. Evaluation metrics

The balanced accuracy measure (BAC) is used to evaluate the model. Balanced accuracy is defined as an average accuracy obtained on either class (male and female) 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.

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

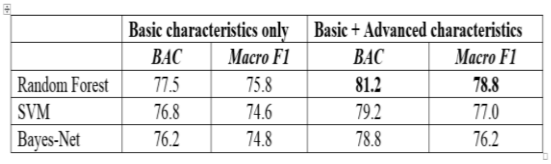
C. Results and Discussion

In order to evaluate the performance of basic and advanced characteristics, we conducted experiments on different sets of characteristics, including basic characteristics only and combination of both types of characteristics. Each set of characteristics was tested using the machine learning methods, namely Random Forest, SVM, and Bayes-Net [18].

The training data and testing datasets are provided separately, each dataset has 10,000 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 and Resampling, [17, 18] we experimented on various combination and found that using cost-sensitive learning solely with cost matrix 1:4 achieved best Macro F1 score (81.4%) but using in combination with resampling filter with cost matrix 1:3 gave best BAC score (81.2%). Table I-II shows the results of our experiments.

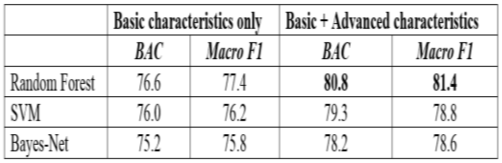
RESULTS OF THE EXPERIMENTS.

TABLE I. RESULTS OF EXPERIMENTS ON COST-SENSITIVE LEARNING WITH SAMPLING METHOD



	Basic characteristics only		Basic + Advanced characteristics	
	BAC	*Macro F1*	*BAC*	*Macro F1*
Random Forest	77.5	75.8	**81.2**	**78.8**
SVM	76.8	74.6	79.2	77.0
Bayes-Net	76.2	74.8	78.8	76.2

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



	Basic characteristics only		Basic + Advanced characteristics	
	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
Bayes-Net	75.2	75.8	78.2	78.6

As the results shown in above TABLE I, when using cost sensitive learning, Random Forest achieved the best results while Bayes Net 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 next TABLE II values 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 characteristics when combining with basic characteristics also remarkably improve prevision result compared with using basic characteristics only. However, in provided datasets, there are many sessions in which users only view one product and the advanced characteristics don't have any effect on these cases.

For the number of characteristics, we selected 2,000 because when we conducted the experiments with different number of characteristics ranging from 500 to 1,500, we found that the prevision result increases and reaches the top at 2.500 characteristics.

V. CONCLUSION

This paper provides our proposed method for previsioning the gender of clients based on product viewing data on ecommerce systems. The approach uses the basic characteristics such as viewing time and duration, individual categories/products along with the advanced characteristics such as categories/products sequences and transfer pairs, taken from the tree-based representation of categories/products list. This characteristic design is work best on the Random Forest technique, a machine learning method with supporting techniques such as Cost-Sensitive Learning and Resampling. This can be easily applied to other datasets because it uses no dataset-specific features. Further this work can be investigated on the characteristic set in future. More type of characteristics can be incurred from the tree-based representation to develop the relation between the products viewed in the same session.

VI. REFERENCES

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

[2] 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

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

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

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

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] J. J. C. Ying, Y. J. Chang, C. M. Huang, and V. S. Tseng, “Demographic prediction based on users mobile behaviours,” In Nokia Mobile Data Challenge, 2012.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.