https://doi.org/10.2352/ISSN.2470-1173.2017.6.MOBMU-298

Demographic Prediction based on Mobile User Data

Podoynitsina L., Romanenko A., Kryzhanovskiy K., Moiseenko A., Samsung R&D Institute Russia, Moscow, Russia

Abstract

Demographic prediction is a very important component to build mobile user profile that can help improve personalized services and targeted advertising. However, demographic information is often unavailable due to user privacy issue. This paper presents technologies and algorithms to build demographic prediction classifiers based on mobile user data such as call logs, app usages, Web data and so on. To associate those data with demographic information, we implemented a system that consists of two parts: mobile application for data collection with web infrastructure for user survey administration (i.e. gender, age, marital status and so on), and classifiers to predict demographic information. In the demographic prediction, we focus on user interest which is semantically extracted from Web data rather than other mobile data. To capture user interest more precisely, advanced topic model called ARTM (Additive Regularization of Topic Models) used. Using user interest as features, the experimental results show our system achieves demographic prediction accuracies on gender, marital status, and age as high as 97%, 94%, and 76%, respectively using deep learning.

Introduction

Mobile phones changed our lives considerably during the past decades. They have evolved from simple communication devices into a primary center for information, entertainment and social interaction. This transition has been encouraged by numerous applications and services supported by modern mobile platforms. At the same time, mobile phones became equipped with various hardware sensors to collect various types of information and software applications to store diverse statistics of user activity.

Our task is to build a demographical model, which will recognize demographic characteristics of user, such as gender, marital status and age. Such demographic information play a crucial role in personalized services and targeted advertising.

Previous research on demographic prediction algorithms has been predominantly focused on separate use of data sources such as web data [1-3], mobile data [3-5] and application data [6]. Our research intends to examine usage of all possible types of available information presented in modern mobile devices and select the best combination of features to improve the prediction accuracy.

However, several issues still remain in the previous work. First problem is that we need to avoid leaking of sensitive information about the user. Since, by its definition, marital status, age and user data tend to be sensitive, it is important to analyze the user behavior on the mobile device. The second problem is that mobile devices have limited resources. Hence, the demographical model needs to be quite lightweight in terms of computational complexity and memory consumption.

We trained our demographic model on server and exported it on mobile device.

We collected an extensive dataset with various available types of features from mobile users with our own application. Users filled form with questions about gender, age, marital status and so on.

We used different types of data for demography prediction like call log, sms log, application usage data and so on. Most of them are easily transformed to be used as features for the demographic predication, except the Web data.

So, the third problem we had to solve was the need to present somehow a massive text data from Web pages in the form of an input feature vector for the demographic model.

Hence, we used advanced NLP technologies based on probabilistic topic modeling algorithm [7] to extract meaningful textual information from the Web data.

We chose to extract user's interests from Web data. On the one side, we obtain a compact representation of a Web user data that can be used for training the demographic classifier. On the other side, user's interests can be directly used by the content provider for better targeting in advertisement services or other interactions with the user.

Hence, it is important to extract information about user interests, for example which books he may read or buy, which sports are interesting for the user or which purchases he could potentially make.

To achieve flexibility of demographic prediction or provide language independence, we decided to use common news categories as a model of user interests. News streams are available in all possible languages of interest. Its categories are also reasonably universal across languages and cultures. It allows us to build multi-lingual topic model.

First we should build and train topic model with classifier of text data to specified categories (interests). The list of wanted categories can be given by content provider. The topic model categorizes the text extracted from Web pages. We build topic model with the Additive Regularization of Topic Models (ARTM) algorithm. Then we extract user interests using trained topic model. ARTM can be used not only for clustering, but for classification for a given list of categories. ARTM is based on generalization of two powerful algorithms: probabilistic latent semantic analysis (PLSA) [9] and latent Dirichlet allocation (LDA) [7]. The additive regularization framework allows imposing additional necessary constraints on the topic model, such as sparseness or desired words distribution.

Second, the demographic model is trained using dataset collected from mobile users.

We extracted features from the collected data with the help of topic model trained on previous step. The demographic model consists of several (in our case 3) demographic classifiers. Demographic classifiers have to predict the age of the user (one of labels "0-18", «19 - 21», «22 - 29, "30+"), gender (male or female) or marital status (married/not married) based on given feature vector.

The fourth problem is that we cannot predict in advance which language will the user use. Hence, it is important that the model needs to be multi-lingual. A speaker of another language may only need to load data for his own language.

ARTM allows inclusion of various types of modalities (translations into different languages, tags, categories, authors, etc.) into one topic model.

We used cross-lingual features to implement a language-independent (multilingual) NLP procedure. The idea of cross-lingual feature generation consists of training one topic model on translations of documents into different languages. The translations are interpreted as modalities, in such a way it is possible to embed texts in different languages into the same space of latent topics.

Topic model

At the first stage, we need to analyze webpages viewed by the user. Our webpage analysis steps consist of preprocessing webpages, probabilistic latent semantic analysis (PLSA), the extension of PLSA with Additive Regularization of Topic Models (ARTM), and document aggregation steps, described in Sections 2.1-2.4, respectively.

Preprocessing

The major source of our observation data is webpages browsed by the user. We preprocess the web pages as follows: remove HTML tags, perform stemming or lemmatization of every word, remove stop words, lowercase all characters and translate webpage content into a target languages. In our system, we have three target languages: Russian, English and Korean. In order to do the translation, in our experiments we used the Yandex machine translation system[1].

Probabilistic Latent Semantic Analysis

In order to model user interests, we need to analyze documents viewed by the user. Topic modeling enables us to assign a set of *topics T* and to use *T* in order to estimate a conditional probability that word *w* appears inside document *d*:

$$p\langle w|d\rangle = \sum_{t\in T} p\langle w|t\rangle p\langle t|d\rangle \qquad (1)$$

where *T* is a set of topics. In line with Probabilistic Latent Semantic Analysis (PLSA) by (Hofmann, 1999)[9], we follow the assumption that all documents in a collection inherit one cluster-specific distribution for every cluster of topic-related words. Our purpose is to assign such topics *T*, which will maximize a functional *L*:

$$L(\varPhi,\varTheta) = ln \prod_{d\in D} \prod_{w\in d} p(w \vee d)^{n_{dw}} \rightarrow \max_{\varPhi,\varTheta} \qquad (2)$$

where n_{dw} denotes the number of times that word *w* is encountered in a document *d*, $\varPhi=(p(w|t))_{W\times T}=(\varphi_{wt})_{W\times T}$ is the matrix of term probabilities for each topic, and $\varTheta=(p(t|d))_{T\times D}=(\theta_{td})_{T\times D}$ is the matrix of topic probabilities for each document. After plugging (1) into (2), we obtain the equations (3), (4):

$$L(\varPhi,\varTheta) = \sum_{d\in D} \sum_{w\in d} n_{dw}\, ln \sum_{t\in T} p(w \vee t)p(t \vee d) \rightarrow \max_{\varPhi,\varTheta} \qquad (3)$$

$$\sum_{w\in W} p\langle w|t\rangle = 1, p\langle w|t\rangle \geq 0; \sum_{t\in T} p\langle t|d\rangle = 1, p\langle t|d\rangle \geq 0; \qquad (4)$$

Additive Regularization of Topic Models

One of the problems in our case is that zero probabilities are not acceptable to the natural logarithm in (3). To overcome this problem, we followed the method by Vorontsov and Potapenko (2015)[8] which they call Additive Regularization of Topic Models (ARTM). First, we added the regularization coefficient $R(\varPhi, \varTheta)$:

$$R(\varPhi,\varTheta) = \sum_{i=1}^{r} \tau_i\, R_i(\varPhi,\varTheta), \tau_i \geq 0 \qquad (5)$$

where τ_i is a regularization coefficient and $R_i(\varPhi, \varTheta)$ is a set of different regularizers. In this work, we used the smoothing regularizers for the both matrices $\varPhi$, $\varTheta$. First, we can define the the Kullback-Leibler divergence as follows:

$$KL(p \vee q) = \sum_{i=1}^{n} p_i\, ln \frac{p_i}{q_i} \qquad (6)$$

The Kullback-Leibler divergence evaluates how well the distribution *p* approximates another distribution *q* in terms of information loss. In order to make the smoothing regularization the values $p(w|t)$ in the matrix $\varPhi$, we need to find such fixed distribution $\beta=(\beta_w)_{w\in W}$ that can approximate $p(w|t)$. Hence, we look for the minimum of KL values:

$$\sum_{t\in T} KL_w\, (\beta_w \vee \phi_{\omega t}) \rightarrow \min_{\varPhi} \qquad (7)$$

Similarly, to make the smoothing regularization of the matrix $\varTheta$, we leverage another fixed distribution $\alpha=(\alpha_t)_{t\in T}$ that can approximate $p(t|d)$:

$$\sum_{d\in D} KL_t\, (\alpha_t \vee \theta_{td}) \rightarrow \min_{\varTheta} \qquad (8)$$

In order to achieve the both minima, we combine the formulas (7) and (8) into a single regularizer $R_s(\varPhi, \varTheta)$:

$$R_s(\varPhi,\varTheta) = \beta_0 \sum_{t\in T} \sum_{w\in W} \beta_\omega\, ln\phi_{\omega t} + \alpha_0 \sum_{d\in D} \sum_{t\in T} \alpha_t\, ln\theta_{td} \rightarrow max \qquad (9)$$

Finally, we combine $L(\varPhi, \varTheta)$ with $R_s(\varPhi, \varTheta)$ in a single formula:

$$L(\varPhi,\varTheta) + R_s(\varPhi,\varTheta) \rightarrow \max_{\varPhi,\varTheta} \qquad (10)$$

We maximize this expression using the EM algorithm by (Dempster et al., 1977)[17].

Document aggregation

After we applied the topical model ARTM, it is possible to describe each topic *t* with the set of its words w using the probabilities $p(w|t)$. We are also able to map each input document *d* into a vector of topics *T* according to probabilities $p(t|d)$. At the next step, we need to aggregate all topical information about the documents $d_{1u}, .., d_{nu}$ viewed by the user *u* into a single vector. In order to do such aggregation, we average the obtained topical vectors $p_u(t_i|d_j)$ as follows:

$$p_u(t_i \vee d) = \frac{1}{N_d} \sum_{j=1}^{N_d} p_u\left(t_i \vee d_{ju}\right) \qquad (11)$$

where N_d is denotes the number of documents viewed by a user *u*, and d_{ju} is the j-th document viewed by the user.

The resulting topic vector (or user interest vector) is used as feature vector for demographic model.

Demographic model

Once the information about topics is aggregated, we need to build a demographic model. Demographic model consist of several demographic classifies. In our case: age classifier, gender classifier and marital status classifier. In this work, we chose to use the deep learning approach using the Veles framework. Each classifier was built with neural network and optimized with genetic algorithm. The architecture of neural network is based on the multi-layer perceptron (Collobert and Bengio, 2004)[18]. We select possible architecture of

[1] translate.yandex.ru

the each neural network and optimal hyper-parameters using the genetic algorithm.

Table I. Parameters adjusted by genetic algorithm

Parameter	Min value	Max Value
#Neurons	8	256
Learning rate	0.0001	0.1
Weights decay	0	0.01
Gradient moment	0	0.95
Weights deviation	0.00001	0.1

We use several hyper-parameters of the neural network architecture: size of minibatch, number of layers, number of neurons in each layer, activation function, dropout, learning rate, weights decay, gradient moment, standard deviation of weights, gradient descent step, regularization coefficients, initial ranges of weights, number of examples per iteration. Genetic algorithm enables us to adjust these hyper-parameters. Also, we used the genetic algorithm to select optimal features in input feature vector and reduce size of input feature vector of demographic model.

When we apply the genetic algorithm, we create a population P with M=75 instances of demographic classifiers with the above-mentioned parameters. Afterwards, we use error backpropagation to train these classifiers. As the result of training, we chose the classifiers with the highest performance in terms of demographical profile prediction. Afterwards, we apply a cross-over operation in order to add new classifiers into the population. Our crossover consists of random substitution of numbers taken from the parameters of two original classifiers, in the case when these parameters are mismatching in the classifiers. For example, if classifier $C1$ contains $n1$=10 neurons in the first layer, and classifier $C2$ contains $n2$=100 neurons, then we may replace its value to 50 in the crossover operation. We use the newly created classifier $C3$=*crossover*($C1$, $C2$) to replace some classifier with the worst performance in the population. We also apply the operation of mutation in order to introduce modifications of best classifiers. We add each classifier with new parameters to the population of classifiers; afterwards we retrain all new classifiers and measure their performance. We continue this process while we can observe the improvement of the classification performance. Finally, we choose the demographic classifier with the best performance in the last population.

Experiments

In this section, we describe our data collection procedure and demographic prediction results.

Data Collection

The information collected from mobile phones of users can be used to predict many types of demographic parameters; however this work mainly focused on gender, marital status and age.

To build robust demographic prediction models, we collected an extensive dataset with various available types of features from mobile users. To accomplish this task, we developed an entire system: 1) Mobile application, which is implemented on Android platform, periodically captures and save user activities on the mobile device with user permission and sends it to a server 2) Server that monitors and controls data collection. More than 500 users have been involved for the data collection which lasts from March 2015 till October 2016. Each user supplied data for at least 10 weeks. To associate the collected data with demographic information, each respondent had to fill a questionnaire with the following questions: date of birth, gender, marital status, household size, job position, work schedule, children and income. Note that the process of data collection, storing and processing was totally anonymized.

The collected mobile user data features were largely divided into three main categories as follows: call+sensors data, application and web data.

Call+sensors data consists of SMS and call log, battery status, cell tower data, light sensor status, location information, magnetic field and Wi-Fi data (some depending on availability).

Application data includes such information as package name, time of installation, price, market name, and category name. Application market-related information is obtained from Google Play store, Amazon App store and Samsung Galaxy Apps store.

Web data is obtained from various browsers (i.e. Google Chrome or Samsung Browser) by using Android platform content provider functions to get history. The history of browsing is then used to get textual (Web page content) information for further analysis by natural language processing (NLP) algorithms.

Demographic Prediction Results

For demographic prediction, we explored different ML approaches and methods such as support vector machines (SVM) [10], neural networks (NNs) [11] and logistic regression [12]. Early accuracy tests (without optimization) has been performed, results presented in table II.

Table II. Tests results (without optimization)

Algorithm	Accuracy (gender) %	Accuracy (marital) %	Accuracy (age) %
Fully connected NNs	84.85	68.66	51.25
Linear SVM	75.76	61.19	42.42
Logistic regression	72.73	52.24	43.94

Through the early tests, we chose the NN approach to build demographic prediction classifier. Currently different deep learning frameworks are available for training NNs: Caffe [13], Torch [14], Theano [15] and others. However, we used our custom deep learning framework (Veles) [16] because it was designed as a very flexible tool in terms of workflow construction, data extraction and preprocessing, visualization, with an additional advantage in the ease of porting of the resulting classifier to mobile devices.

We implemented genetic algorithm to find the optimal NN configuration. Such parameters of NNs as topology (number of neurons in layers) and hyper-parameters (learning rate, weights decay, etc.) had been optimized. The number of topic features had also been optimized. For instance, the initial number of ARTM features was decreased from 495 to 170. The demographic prediction accuracies using topic features generated from ARTM and LDA are shown in Table III. Our final experimental results shows our system achieves demographic prediction accuracies on gender, marital status, and age as high as 97%, 94%, and 76%, respectively.

Table III. A comparison of accuracy b/w ARTM and LDA

Task	Demographic Prediction Accuracy (%)	
	ARTM	LDA
Gender	93.7	88.9
Marital status	87.3	79.4

Age	62.9	61.3
Gradient moment	0	0.95
Weights deviation	0.00001	0.1

Conclusion

This work describes novel approaches for demographic prediction using mobile user data. To achieve highly accurate result, we examine usage of all available in modern mobile phone sources of information.

To find the best solution, we build a framework from data collection campaign to final model fine-tuning by testing different algorithms with various sources. NN approach based on user interest captured by ARTM is the best candidate for the demographic predication.

The obtained models can be useful to infer user demographics in content targeting or mobile phone personalization.

References

[1] Hu J. et al. Demographic prediction based on user`s browsing behavior. Microsoft research Asia. May 8-12, Banif, Alberta, Canada (2007)

[2] Kabbur, S., Han, E., Karypis G. Content-based methods for predicting web-site demographic attributes. Department of Computer Science and Engineering University of Minnesota. TR 10-021. (2010)

[3] Laurila J.K. et al. From big smartphone data to worldwide research: The Mobile Data Challenge. Pervasive and mobile computing. Vol 9. 752–771 (2013)

[4] Zhong E. et al. User demographic prediction based on mobile data. Pervasive and mobile computing. Vol 9. 823–837 (2013)

[5] Dong Y. et al. Inferring user demographics and social strategies in mobile social networks. ACM digital library. http://dl.acm.org/citation.cfm?doid=2623330.2623703 (2014)

[6] Seneviratne S. et al. Predicting User Traits From a Snapshot of Apps Installed on a Smartphone. Mobile Computing and Communications Review. Vol 18. (2014)

[7] Blei, D.M. Probabilistic topic models. Communications of the ACM. 55(4) 77–84. (2012)

[8] Vorontsov K., Potapenko A. Additive Regularization of Topic Models Machine Learning Journal, Special Issue “Data Analysis and Intelligent Optimization”, Springer (2014)

[9] Hofmann T. Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM 50–57. (1999)

[10] C. Cortes and V. Vapnik. Support-vector networks. Machine Learning, 20(3):273–297. (1995)

[11] C. Bishop. Neural networks for pattern recognition. Clarendon Press. Oxford. (1995)

[12] C. Bishop. Pattern recognition and machine learning. Springer Science + Business Media, LLC. (2006)

[13] Deep learning framework. http://caffe.berkeleyvision.org/. (2016)

[14] Scientific computation for LuaJIT. http://torch.ch/. (2016)

[15] Theano 0.8.0. http://deeplearning.net/software/theano/. (2016)

[16] Distributed machine learning platform. https://github.com/Samsung/veles. (2016)

[17] A.Dempster, N. Laird and D. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society, Series B, 1-38.

[18] R. Collobert and S. Bengio. 2004. Links between perceptrons, MLs and SVMs. In ICML.

Author Biography

Lyubov Podoynitsina graduated from Lomonosov Moscow State University, specialist in Mathematics in 2013. Successfully passed two-month internship in Samsung and was hired as a junior engineer at Samsung R&D Institute, Russia in September 2013. Promoted to engineer in March, 2015. Working on machine learning problems, user profiling, traffic recognition, gestures recognition, brands recognition and so on.

Alexander Romanenko is PhD Student at Moscow Institute of Physics and Technology (State University). Bachelor (2012) and Master (2014) of Physico-Mathematical Sciences at MIPT(SU) (Department of Applied Mathematics and Control), Diplomas with Honours. Joined Samsung R&D Institute Russia in 2012. Main fields of interest: Machine Learning and Data Analysis, NLP, Statistical Analysis of large Text Collections, (Probabilistic) Topic Modeling

Konstantin Kryzhanovskiy received his MS degree in Computer Science from Moscow Engineering Physics Institute (State University) (2000). Works in Samsung R&D Institute Russia since 2011. Research interests are image enhancement and segmentation, features extraction and pattern recognition problems.

Andrey Moiseenko graduated from Saint-Petersburg State Polytechnical University in 2002 with engineer's diploma in Radiophysics and Electronic. Since 2003 works as software engineer. Joined Samsung R&D Institute Russia in 2012. Main research areas include image processing algorithms, development of audio/video codecs on DSP platforms, data analysis.