移动用户人口统计信息预测

王亦雷[1]，嵇智源[2]，夏 勇[1]，秦 臻[1]，程红蓉[3]

(1. 电子科技大学信息与软件工程学院 成都 610054; 2. 科技部高技术研究发展中心 北京 海淀区 100044;
3. 电子科技大学计算科学与工程学院 成都 611731)

【摘要】提出了一种基于支持向量机的预测方法，通过分析智能手机应用的使用情况，预测用户的人口统计信息。手机使用行为数据约为5万智能手机用户在3个月期间使用手机应用产生的网络日志文件，包括179 954 181条日志记录。通过对日志记录的主题进行分析，可将179 954 181条日志记录匹配到266个不同的主题。在此基础上，通过将每个用户的人口统计信息与该用户对266个不同主题的访问权重进行关联，可构建训练数据，并代入支持向量机模型进行计算。实验结果表明该方法对用户的性别和年龄预测能够取得良好的预测结果。

关 键 词 人口统计信息; 预测; 智能手机应用; 支持向量机
中图分类号 TP393 **文献标志码** A doi:10.3969/j.issn.1001-0548.2015.06.021

Demographic Information Prediction for Mobile Users

WANG Yi-lei[1], JI Zhi-yuan[2], XIA Yong[1], QIN Zhen[1], and CHENG Hong-rong[3]

(1. School of Information and Software Engineering, University of Electronic Science and Technology of China Chengdu 610054;
2. High Technology Research and Development Center, Ministry of Science and Technology Haidian Beijing 100044;
3. School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 611731)

Abstract A support-vector-machine-based predicting method is presented to predict users' demographic information by analyzing the usage of the applications in the smartphones. The smartphone usage data considered in this paper is a network log file, which records smartphone applications usage of 50 000 smartphone users for three months, including 179 954 181 entries. By analyzing the topic of each entry, the 179 954 181 entries can be matched with 266 distinct topics. Based on this result, by correlating the users' demographic information with their query weight of such 266 distinct topics, a training data can be constructed and imported to support vector machine model for computation. The results of experiments show that the method proposed in this paper can well predict uses' gender and age.

Key words demographic information; prediction; smartphone application; support vector machine

随着移动互联网的发展，许多互联网公司越来越关心用户的基础属性信息，包括性别、年龄、收入及文化水平等，以便于提供更好的个性化服务。如Google公司提供的个性化搜索服务，结合了用户的地理位置信息返回相应的搜索结果[1]；亚马逊购物网站则根据用户的浏览和购买记录，向用户推荐相应的商品以促进用户消费。与此同时，定制广告投放也是一种越来越流行的个性化服务，定制广告投放是指根据用户的兴趣爱好投放相应的广告[2]。最近的研究表明定制广告投放可以获得比普通广告投放更好的宣传效果[3]。

在个性化服务和定制广告投放业务中，用户的浏览记录、搜索兴趣、地理位置信息和人口统计信息等个人信息扮演着重要的角色。其中，用户的人口统计信息(如性别、年龄、收入和文化程度等)尤为重要。然而，人口统计信息是用户比较敏感的隐私信息，用户不愿意公开这类隐私属性，这类信息不易获取。

尽管如此，近年来很多学者通过分析用户的行为数据(如博客、照片、社交网站状态、心情评论等)获取用户的人口统计信息。文献[4]指出通过研究用户的书写和说话方式可以预测出用户的人口统计信息；文献[5]的研究表明通过分析博客内容可以预测博客作者的性别；文献[6]通过研究Twitter用户在

收稿日期：2014－04－14；修回日期：2014－11－28
基金项目：国家自然科学基金(61133016, 61300191, 61370026)；教育部-中国移动科研基金(MCM20121041)；四川省科技支撑计划(2014GZ0106)；中央高校基金(ZYGX2013J003)
作者简介：王亦雷(1985－)，男，博士生，主要从事移动互联网、数据挖掘、信息安全等方面的研究.

Twitter上发表的内容来预知用户的性别。此外，还有一些学者通过分析用户的搜索历史记录和浏览历史记录等Internet行为数据，分析用户的人口统计信息。文献[7]研究不同性别和年龄的人搜索行为之间的差异性，并且发现搜索引擎用户的基础属性分布和美国的人口分布相符；文献[8]的另一项研究表明用户搜索的内容和用户的性别、年龄等是相关联的；文献[9]通过分析用户浏览网页的历史记录判别出用户的性别和年龄；文献[10]通过分析用户浏览网页的内容和关键字预测用户的性别和年龄。

随着移动互联网的发展，智能手机应用成了人们生活中必不可少的重要组成部分。但是由于用户的个体需求和兴趣的差异，每个用户智能手机上安装的应用有所差别。如，男性更偏好运动类的手机应用，而女性则比较喜欢时尚类的手机应用。即便对于相同的应用，不同的用户也会有不同的使用偏好。如，对于一个网络视频应用，成年人更倾向于观看时政新闻，而儿童则更倾向于观看娱乐节目。由于智能手机和用户是紧密相关的，分析手机的使用行为使得预测用户的人口统计信息成为可能，本文将尝试通过分析用户智能手机上应用的使用情况，进而预测用户的性别和年龄。

1 问题定义和数据说明

本文旨在通过分析一定数量的人口统计信息已知的用户的智能手机应用情况，结合部分人口统计信息未知的用户的智能手机应用情况，对其他用户的人口统计信息进行预测。

本文着重关注用户的性别和年龄。用户的性别预测被定义为将用户分类为男性或者女性的一个二分类问题；用户的年龄预测被定义为一个多分类问题，分类类别如表1所示。

表1 年龄分组

分组	区间
少年	<18
青少年	18～24
青年	25～44
中年	45～60
老年	>60

本文的数据集是：国内一家网络运营商提供的近5万智能手机用户在2013年10月–2013年12月3个月期间使用智能手机应用产生的网络日志文件。当智能手机应用向Internet获取资源时，则产生一条日志，记录在日志文件中。数据集中一共有179 954 181条日志记录，每一条日志记录由用户的ID、应用名称和相应的网络资源组成。数据集中用户的性别和年龄分布如表2所示。

表2 用户基础属性分布

年龄分布/%	性别分布		总计/%
	男性/%	女性/%	
少年	0.38	0.26	0.64
青少年	6.02	4.23	10.25
青年	38.93	22.98	61.91
中年	14.95	7.15	22.11
老年	3.65	1.44	5.09
总计/%	63.94	36.06	100.00

对于上述日志文件，本文中采用正则表达式将相应的网络资源匹配到相应的主题(如，将德甲归类到运动\足球\欧洲足球\德甲)。每一条记录都映射到一个主题。本文将每个主题定义为用户的一个兴趣。通过匹配，最终本文将日志文件中所有的记录匹配到266个主题(兴趣)。

这样，本文中使用的数据集则可以定义为一个有权有向偶图 $G=(V,E)$，V 是顶点的集合，E 是边的集合。顶点集 V 中的一个顶点代表某个用户或者某个用户的一项兴趣类别；边集 E 中的一条边代表某个用户对某个兴趣类别的偏好程度。进一步，顶点集 V 可划分为两个子集合 $U=\{u_1,u_2,\cdots,u_m\}$ 和 $C=\{c_1,c_2,\cdots,c_n\}$，其中子集U为用户集，子集C为用户的兴趣集。如果用邻接矩阵$\boldsymbol{R}$表示有权有向图 G，那么邻接矩阵$\boldsymbol{R}$中的元素 r_{ij} 代表用户 i 对兴趣类别 j 的偏好程度。本文中取用户 i 对兴趣类别 j 的请求次数来评估用户对其的偏好程度。将用户的性别、年龄和邻接矩阵$\boldsymbol{R}$相结合，可以统计出具有不同性别、年龄的用户兴趣分布，其分布如图1所示。

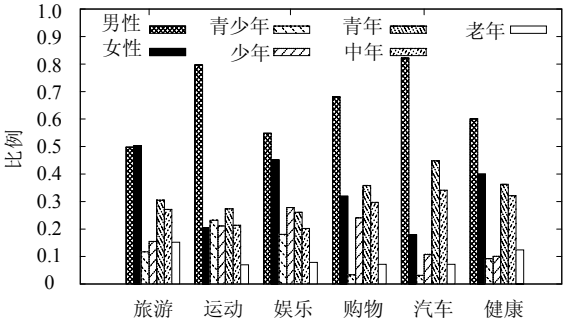

图1 具有不同性别、年龄的用户兴趣分布

2　方法说明

2.1　数据分析

性别和年龄相似的用户可能有相似的兴趣爱好，有相似兴趣爱好的用户也很可能具有相似的性别和年龄。根据该假设，若采用一种直接的方法，可以用协同过滤[11-12]的方法预测用户的性别和年龄，但是因为数据(邻接矩阵$\boldsymbol{R}$)的稀疏，而协同过滤对数据的稀疏性很敏感[13]，如果直接采用协同过滤方法进行预测，会引入很多的噪声，对预测效果造成不良影响。

为了解决该问题，本文将采用奇异值分解(singular value decomposition)[14-16]技术对邻接矩阵$\boldsymbol{R}$进行预处理。经过奇异值分解之后，可以得到相互正交的向量，避免原始数据(邻接矩阵$\boldsymbol{R}$)行列向量之间的干扰,进而可以更好地挖掘数据间的隐性关系[14]。

2.2　方法步骤

本文首先用余弦相似性计算用户之间的相似度，得到用户的相似性矩阵；然后利用SVD技术[14]分解用户的相似性矩阵，得到用户的隐性反馈矩阵；最后将隐性反馈矩阵作为特征向量输入高斯核的支持向量机[17]预测用户的性别和年龄。

根据邻接矩阵$\boldsymbol{R}$，采用余弦相似性计算出用户(子集U中的元素)之间的相似性，计算公式为：

$$\mathrm{sim}(R_i,R_j)=\frac{\sum_{k=1}^{n}R_{ik}R_{jk}}{\sqrt{\sum_{k=1}^{n}R_{ik}^2\sum_{k=1}^{n}R_{jk}^2}} \tag{1}$$

式中，R_i是用户i的兴趣偏好向程度。计算后，可得到用户的相似性矩阵$\boldsymbol{S}$。采用SVD将用户的相似性矩阵$\boldsymbol{S}$分解为两个低维矩阵相乘：

$$\hat{\boldsymbol{S}}=\boldsymbol{P}\times\boldsymbol{P}^{\mathrm{T}} \tag{2}$$

式中，$\boldsymbol{P}\in\boldsymbol{R}^{m\times k}$是降维后的矩阵。那么用户$i$和用户$j$之间的相似程度可以通过如下公式计算：

$$\hat{s}_{ij}=\sum_{k}p_{ik}p_{jk} \tag{3}$$

式中，$p_{ik}=P(i,k)$，$p_{jk}=P(j,k)$。通过训练，利用最小均方根误差(root mean square error)学习$\boldsymbol{P}$矩阵。同时，为了防止过拟合，在损失函数中加入过拟合项，其定义为：

$$C(p)=\sum\left(s_{ij}-\sum_{f=1}^{k}p_{if}p_{jf}\right)^2+\lambda(\|p_i\|^2+\|p_j\|^2) \tag{4}$$

为最小化损失函数，采用随机梯度下降算法[18]求解参数$\boldsymbol{P}$。根据随机梯度下降算法，先对式(4)中的参数p_{if}和p_{jf}求偏导数，求解公式为：

$$\begin{cases}\dfrac{\partial C(p)}{\partial p_{if}}=-2p_{jf}(s_{ij}-\sum_{f=1}^{k}p_{if}p_{jf})+2\lambda p_{if}\\ \dfrac{\partial C(p)}{\partial p_{jf}}=-2p_{if}(s_{ij}-\sum_{f=1}^{k}p_{if}p_{jf})+2\lambda p_{jf}\end{cases} \tag{5}$$

然后，需要将参数沿着梯度的方向更新，递推公式为：

$$\begin{cases}\mathrm{err}=s_{ij}-\sum_{f=1}^{k}p_{if}p_{jf}\\ p_{if}=p_{if}+\alpha(p_{jf}\times\mathrm{err}-\lambda p_{if})\\ p_{jf}=p_{jf}+\alpha(p_{if}\times\mathrm{err}-\lambda p_{jf})\end{cases} \tag{6}$$

当误差err小于某一个设定的阈值时停止迭代。在迭代更新学习速率α的取值需要通过反复实验获得。如后面实验所示，在实验开始时需要对矩阵$\boldsymbol{P}$进行初始化，学习速率α在每一步学习时需要进行衰减。

在学习完成之后可以得到用户的隐私反馈矩阵$\boldsymbol{P}$，将用户隐私反馈矩阵$\boldsymbol{P}$和用户的基础属性相结合，$\boldsymbol{P}$作为特征矩阵，采用高斯核的SVM分类，对用户的性别和年龄进行预测。

3　实验结果

3.1　评价指标

本文中采用准确率(Acc)、精确率(Prec)、召回率(Rec)和F_1值[9]作为评价指标。准确率(Acc)定义为正确预测的用户数和实际用户的总人数的比值，精确率(Prec)定义为正确预测为某类的人数和预测为该类的人数的比值，召回率(Rec)定义为正确预测某类人数和该类实际人数的比值，F_1值是精确率和召回率的权衡，计算公式为:

$$F_1=\frac{2\mathrm{Prec}\times\mathrm{Rec}}{\mathrm{Prec}+\mathrm{Rec}} \tag{7}$$

因为预测有多个类别，所以本文中采用宏观的F_1值作为评价指标。

3.2　实　验

根据本文提出的方法，首先初始化矩阵$\boldsymbol{P}$。对矩阵$\boldsymbol{P}$的初始化有多种方法，一般是将$\boldsymbol{P}$用随机数填充[9]。在实验中，则是用和sqrt(k)成反比的高斯分布随机数初始化矩阵$\boldsymbol{P}$。参数λ和α以最小化损失函数$C(p)$为目标，可通过交叉验证得出。本文通过反复实验得出α=0.01，$\lambda=0.001$，且α在每一步学习之后自乘0.9衰减。

在SVD分解中，SVD的维度k是一个重要的参

数，通过实验研究k对预测结果的影响。最小化损失函数$C(p)$，设定维度k从5～100逐步变化。对于每一个维度k，设定SVD迭代次数从1～150变化，进行反复迭代学习，得到每个维度下的最优迭代次数，从而得到隐私反馈矩阵$\boldsymbol{P}$。以性别预测为例，SVD维度k对预测结果的影响效果如图2所示。

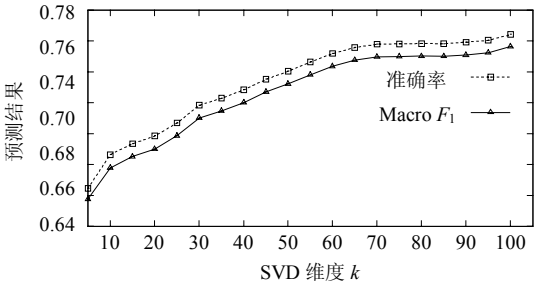

图2 SVD维度k对预测结果的影响

从图2可以看出，随着维度k的增加，预测结果的准确率和F_1值都有所提升，当k值达到70时，得到一个较稳定的预测结果，准确率为75.79%和F_1值为74.96%。在后面的分类实验中将$\boldsymbol{P}$矩阵作为特征向量，采用高斯核的支持向量机分类方法分类预测用户的基础属性。对用户的性别和年龄分别训练分类模型，实验采用10倍交叉验证法，对性别和年龄进行预测。

用户的性别和年龄的预测结果如表3所示。用户的年龄分类预测是一个五分类问题，预测结果达到准确率57.14%和F_1值52.52%；对于用户的性别分类这样的二分类问题，预测效果更佳，达到准确率76.29%和F_1值75.21%。

表3 用户基础属性预测结果

分组	Prec /%	Rec /%	Micro F_1 /%	Macro F_1 /%	Acc /%
男性	75.72	85.68	80.39	75.21	76.29
女性	77.32	63.98	70.02		
少年	53.10	32.50	40.33		
青少年	53.39	62.33	57.52		
青年	54.74	45.31	49.58	52.52	57.14
中年	67.36	62.44	64.80		
老年	42.70	61.42	50.38		

4 结 束 语

本文提出的预测方法可以根据移动用户智能手机应用的使用情况，预测用户性别、年龄等用户隐私属性。该预测方法主要包含3个步骤：1) 将智能手机用户的手机应用每条日志记录匹配相应的主题，从而得到一个关联用户和兴趣类型的邻接矩阵；2) 结合用户的兴趣偏好计算用户的相似相关性，得到用户的相关性矩阵，再采用SVD分解技术，分解用户的相关性矩阵以得到用户的隐性反馈矩阵；3) 将用户的隐性反馈矩阵作为用户的特征，采用高斯核的支持向量机分类器分别训练用户的性别和年龄的分类模型。基于运营商的现实数据，采用交叉验证的实验结果显示本文的方法对用户的性别、年龄预测能够取得很好的分类预测效果，对用户的性别的预测能够达到76.29%的准确率和75.21%的F_1值，对用户的年龄预测能够达到准确率57.14%和52.52%的F_1值。

参 考 文 献

[1] HANNAK A, SAPIEZYNSKI P, MOLAVI K A, et al. Measuring personalization of web search[C]//Proceedings of the 22nd International Conference on World Wide Web. Switzerland: International World Wide Web Conferences Steering Committee, 2013: 527-538.

[2] SMIT E G, VAN N G, VOORVELD H A M. Understanding online behavioural advertising: User knowledge, privacy concerns and online coping behaviour in Europe[J]. Computers in Human Behavior, 2014, 32: 15-22.

[3] JANSEN B J, MOORE K, CARMAN S. Evaluating the performance of demographic targeting using gender in sponsored search[J]. Information Processing & Management, 2013, 49(1): 286-302.

[4] GARERA N, YAROWSKY D. Modeling latent biographic attributes in conversational genres[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Stroudsburg: Association for Computational Linguistics, 2009: 710-718.

[5] YAN X, YAN L. Gender classification of weblog authors[C]//AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. California: AAAI, 2006: 228-230.

[6] BURGER J D, HENDERSON J, KIM G, et al. Discriminating gender on twitter[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2011: 1301-1309.

[7] WEBER I, CASTILLO C. The demographics of web search[C]//Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2010: 523-530.

[8] WEBER I, JAIMES A. Demographic information flows[C]//Proceedings of the 19th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2010: 1521-1524.

[9] HU J, ZENG H J, LI H, et al. Demographic prediction based on user's browsing behavior[C]//Proceedings of the 16th International Conference on World Wide Web. New York: ACM Press, 2007: 151-160.

(下转第933页)

[4] 姜志宏. 大规模P2PTV系统测量与建模研究[D]. 长沙: 国防科学技术大学, 2011.
JIANG Zhi-hong. Research on modeling and measurement of large scale P2P TV systems[D]. Changsha: National University of Defense Technology, 2011.
[5] 徐恪, 张赛, 陈昊, 等. 在线社会网络的测量与分析[J].计算机学报, 2014, 37(1): 165-188.
XU Ke, ZHANG Sai, CHEN Hao, et al. Measurement and analysis of online social networks[J]. Chinese Journal of Computers, 2014, 37(1): 165-188.
[6] MISLOVE A, MARCON M, GUMMADI K P, et al. Measurement and analysis of online social networks[C]// Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. [s.l.]: ACM, 2007: 29-42.
[7] WILSON C, BOE B, SALA A, et al. User interactions in social networks and their implications[C]//Proceedings of the 4th ACM European Conference on Computer Systems. [s.l.]: ACM, 2009: 205-218.
[8] MATEI R, IAMNITCHI A, FOSTER I. Mapping the Gnutella network[J]. Internet Computing, 2002, 6(1): 50-57.
[9] 汪小帆, 李翔, 陈关荣. 复杂网络理论及其应用[M]. 北京: 清华大学出版社有限公司, 2006.
WANG Xiao-fan, LI Xiang, CHEN Guan-rong. Complex networks theory and its application[M]. Beijing: Tsinghua university press co, LTD, 2006.
[10] NEWMAN, MARK E J. The structure and function of complex networks[J]. SIAM Review, 2003, 45(2): 167-256.
[11] JIANG J, WILSON C, WANG X, et al. Understanding latent interactions in online social networks[J]. ACM Transactions on the Web (TWEB), 2013, 7(4): 18.

编 辑 蒋 晓

(上接第920页)

[10] KABBUR S, HAN E H, KARYPIS G. Content-based methods for predicting web-site demographic attributes [C]//2010 IEEE 10th International Conference on Data Mining (ICDM). Sydney: IEEE Press, 2010: 863-868.
[11] BREESE J S, HECKERMAN D, KADIE C. Empirical analysis of predictive algorithms for collaborative filtering[C]//Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc, 1998: 43-52.
[12] SU X, KHOSHGOFTAAR T M. A survey of collaborative filtering techniques[EB/OL]. [2014-01-15]. http://www.hindawi.com/journals/aai/2009/4214251.
[13] SARWAR B, KARYPIS G, KONSTAN J, et al. Application of dimensionality reduction in recommender system-a case study[R]. Minneapolis: Dept of Computer Science Univ of Minnesota, 2000.
[14] KOREN Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2008: 426-434.
[15] PRYOR M H. The effects of singular value decomposition on collaborative filtering[R]. Hanover: Dartmouth College, 1998.
[16] GOLUB G H, VAN LOAN C F. Matrix computations[M]. Maryland: Johns Hopkins University Press, 2012.
[17] JOACHIMS T. Making large scale SVM learning practical[R]. Dortmund: Universitat Dortmund,1999.
[18] LECHEVALLIER Y, SAPORTA G. Blum MGB choosing the summary statistics and the acceptance rate in approximate Bayesian computation[C]//Proceedings of Computational Statistics. Herdelberg: Springer, Physica Verlag, 2010: 47-56.

编 辑 蒋 晓