浙江科技学院学报，第 31 卷第 6 期，2019 年 12 月
Journal of Zhejiang University of Science and Technology
Vol. 31 No. 6, Dec. 2019

doi：10.3969/j.issn.1671-8798.2019.06.006

基于社交网络 LinkedIn 的用户年龄估计

师磊磊[1]，万　健[1]，司华友[2]，陈彬彬[2]
（1.浙江科技学院 信息与电子工程学院，杭州 310023；2.杭州电子科技大学 计算机学院，杭州 310018）

摘　要：基于职业社交网络 LinkedIn 的用户年龄估计方法的研究，对用户的职业发展趋势、职业适应性分析，以及设计更合理的职业推荐系统具有积极的意义。通过挖掘分析用户的个人资料，设计年龄估计模型（age estimation method，AEM），描述年龄与教育和工作经历的关系。结果表明，在社交网络 LinkedIn 中，AEM 较人脸识别年龄估计方法有更高的准确性，体现 AEM 具有一定的研究价值。

关键词：年龄估计模型；人脸识别；LinkedIn
中图分类号：TP391.4　　**文献标志码**：A　　**文章编号**：1671-8798(2019)06-0464-06

Age estimation of users in social network LinkedIn

SHI Leilei[1], WAN Jian[1], SI Huayou[2], CHEN Binbin[2]
(1. School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, Zhejiang, China; 2. School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, Zhejiang, China)

Abstract: The study of user age estimation method based on the professional social network LinkedIn, is of positive significance for career development trend analysis, user's career adaptability analysis and designing a more reasonable career recommendation system. By mining and analyzing the user's profile, the age estimation model(AEM) was designed to describe the relationship between age and education, work experience. The results show that in the social network LinkedIn, AEM has better accuracy than the face recognition age estimation method, revealing certain research value of AEM.

Keywords: age estimation model(AEM); face recognition; LinkedIn

年龄作为社交网络中关键的用户属性之一，在研究互联网用户行为分析方面有着重要的地位，并且在推荐系统和兴趣分析等研究领域也发挥着重要的作用。近些年国内外在年龄估计方面已经进行了许

收稿日期：2019-05-11
基金项目：浙江省大学生科技创新活动计划（新苗人才计划）（F518310J34）
通信作者：万　健（1969—　），男，福建省泉州人，教授，博士，主要从事云计算及大数据研究。E-mail：wanjian@zust.edu.cn。

多研究，大部分的研究成果都集中在深度学习、人脸识别领域。深度卷积神经网络利用弱标记数据可以获得不错的人脸识别性能[1]。通过改进优化神经网络，x 向量神经网络结构[2]、Ranking-CNN(convolutional neural network)框架[3]、不使用面部标志方法[4]、深度年龄分布学习方法[5]及少数据集也适用的卷积神经网络[6]等方法在年龄估计中均有较高的准确率。建立完善的人脸数据库，并在此基础上进行分类能得到偏差较小的年龄估计结果，如吴仰波等[7]建立的亚洲人脸数据库。针对基于人脸图像估计目标年龄的问题，结合 CNN 和最小二乘支持向量机(least squares support vector machine，LSSVM)的年龄估计方法也得到了较好的结果[8]。针对年龄估计模型本身的研究提高了估计的准确率，例如基于区域的动态权值年龄估计模型[9]、多任务估计模型[10]、两层年龄估计模型[11]等。张珂等[12-13]利用多级残差网络和自适应多元回归的方法提高了年龄估计的正确率。除了面部识别进行人类年龄估计外，Mccomb等[14-17]通过分析和处理说话者的声音，估计其性别和年龄。近些年，根据社交网络中用户提供的文本信息进行年龄估计方面的研究不多。陈敬等[18-19]通过分析用户在网络中发布的文本信息或者用户的位置信息、通话频率、性别等信息来推断用户的年龄和性别。以上这些方法应用于社交网络尚存在一些问题，如用户在社交网络上照片的时效性及真实性(非本人或图片失真等)使得人脸识别年龄估计的方法误差较大；在一般的社交网络中用户的语音资料极少，并且同样存在时效性和真实性(非本人或处理过等)的问题；基于用户日常发布的文本挖掘分析的年龄估计，局限于一定的社交应用场景和社交方式。而在职业社交网络中，用户几乎不发布自己的日常信息而是不断完善自己的简历信息(包括教育经历、工作经历、掌握的技能等)，因此，我们提出根据用户的教育经历和工作经历来估计用户年龄的模型 AEM(age estimation model)。

1　数据分析

1.1　数据采集

通过高效的分布式爬虫系统，使用 LinkedIn 官方 API 获取用户的简历数据。用户的简历数据主要包括教育经历、工作经历、项目经历、职业技能及职业技能获得的点赞数等信息。我们获得 64 442份用户简历数据，数据清洗后最终得到 41 366 份同时包含教育经历和工作经历的简历数据。本研究主要关注用户的教育经历和工作经历，而用户的教育经历和工作经历通常由多个时间段构成，为了更方便地处理数据及提高数据的可读性，将数据处理为 json 文件并保存到 MongoDB 数据库中。

LinkedIn 中不提供用户的真实年龄，但真实年龄对定义模型、验证模型有着至关重要的意义，因此我们采用小组多人独立估计的方式进行人工分析标注用户年龄。小组由 5 人构成，首先 5 人分别读取分析用户的简历数据，独立地给出年龄估计值；然后将 5 个年龄估计值的均值作为用户的“真实年龄”。按照以上人工标注的方法，我们在 41 336 份数据中标注了 20 000 份数据。

1.2　数据优先级分析

分析大量的用户简历数据后发现，不同教育阶段对应的年龄段相对固定，并且教育阶段越低对应的年龄段越固定，如大多数人是6岁上小学，12岁上初中。因为在年龄较小时学习阶段是在监护人的安排下有序进行，相同教育阶段年龄差异较小。随着年龄的增长，人的思想逐渐成熟，兴趣和追求的目标也不尽相同，这大大增加了在相同教育阶段中存在年龄差异的概率。相比于教育经历，工作经历对应的年龄不确定性更大，但通常认为人们参加工作的年龄应该是在最后一个教育阶段结束后。由于存在先工作再教育的情况，故在第 2 章作进一步的比较分析。根据以上分析，为了更准确地进行年龄估计，定义数据选用顺序的优先规则：1)教育经历优先工作经历；2)教育经历越早优先级越高；3)工作经历越早优先级越高。表 1 给

表 1　教育阶段和对应的年龄

Table 1　Education stage and corresponding age

阶段	起始年龄/岁	优先顺序
小学	6	1
初中	12	2
高中	15	3
大学	18	4
第一份工作	22	5

出人们在不同教育阶段数据的优先使用顺序，以及对应的起始受教育年龄。本研究将每个阶段对应的起始年龄作为 5 人小组人工标注的参考。

2 年龄估计模型

2.1 模型定义

定义 $f(S,T)$ 来描述年龄与教育和工作经验之间的关系。假设 $S=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_g\}$，$T=\{\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_g\}$ 为所有用户的教育数据集合和工作数据集合，其中 g 为用户总数。$\boldsymbol{s}_i=(s_{i1},s_{i2},\cdots,s_{in})$，$\boldsymbol{t}_i=(t_{i1},t_{i2},\cdots,t_{im})$，$n$ 和 m 分别为用户的教育阶段数和工作经历数，s_{ij} 为第 $i(i\in\{1,\cdots,g\})$ 个用户的教育经历中第 $j(j\in\{1,\cdots,n\})$ 个阶段，t_{ik} 为第 $i(i\in\{1,\cdots,g\})$ 个用户的工作经历中第 $k(k\in\{1,\cdots,m\})$ 个阶段。

为了使年龄估计的模型能更好地描述年龄与教育和工作经验之间的关系，添加偏置量 w。假定年龄估计值用 y 来表示，那么年龄估计模型可以定义如下：

$$y=f(S,T)+w。$$

事实上我们在进行年龄估计时，会根据 **1.2** 的数据选择优先级，选取一段用户优先级最高的教育或者工作经历，并得到该段经历的起始时间，比如 1993(年)等。根据起始时间定义年龄估计函数，令函数值不断逼近真实年龄，最终获得最优参数，完成年龄估计模型的定义。

根据数据优先级，选取教育阶段的起始时间为 p，选取工作阶段的起始时间为 q，那么教育经历估计模型和工作经历估计模型分别定义如下：

$$f(S)=\alpha p+\beta; \tag{1}$$

$$f(T)=\lambda q+\eta。 \tag{2}$$

式(1)～(2)中：α、β、λ、η 均为常数。

如果将人工标注的年龄作为用户“真实年龄”，用 y' 表示，估计误差用 e 表示，那么根据教育经历和工作经历，第 i 个用户的估计误差 e_i 分别为 $\alpha p_i+\beta-y_i'$ 和 $\lambda q_i+\eta-y_i'$。

定义估计误差平方和为 φ，那么教育数据估计模型的误差平方和可以表示如下：

$$\varphi=\sum_{i=1}^{g}(\alpha p_i+\beta-y_i')^2。 \tag{3}$$

分别对式(3)的 α、β 求偏导，且使偏导等于 0，可得：

$$\begin{cases} g\beta+\left(\sum\limits_{i=1}^{g}P_i\right)\alpha=\sum\limits_{i=1}^{g}y_i'; \\ \left(\sum\limits_{i=1}^{g}P_i\right)\beta+\sum\limits_{i=1}^{g}(P_i^2)\alpha=\sum\limits_{i=1}^{g}y_iP_i。 \end{cases} \tag{4}$$

由式(4)解得 α、β，进而得到 $f(S)$。同理可得 λ、η、$f(T)$。用得到的模型估计用户年龄，定义估计年龄和真实年龄的差值为 ε，则偏置量 w 可以表示如下：

$$w=\frac{1}{g}\sum_{i=1}^{g}\varepsilon_i。$$

因此可以得到最终的年龄估计模型 AEM，包括教育经历年龄估计模型 AEM1 和工作经历年龄估计模型 AEM2：

$$y=\alpha p+\beta+w_1; \tag{5}$$

$$y=\lambda q+\eta+w_2。 \tag{6}$$

式(5)～(6)中：w_1 和 w_2 分别是教育经历年龄估计和工作经历年龄估计的偏置量。

2.2 年龄估计过程

首先根据用户的简历信息，提取教育经历和第一份工作的入职时间。然后遵循优先规则选择数据和年龄估计模型；最后使用年龄估计模型计算用户的年龄。年龄估计过程如图 1 所示。

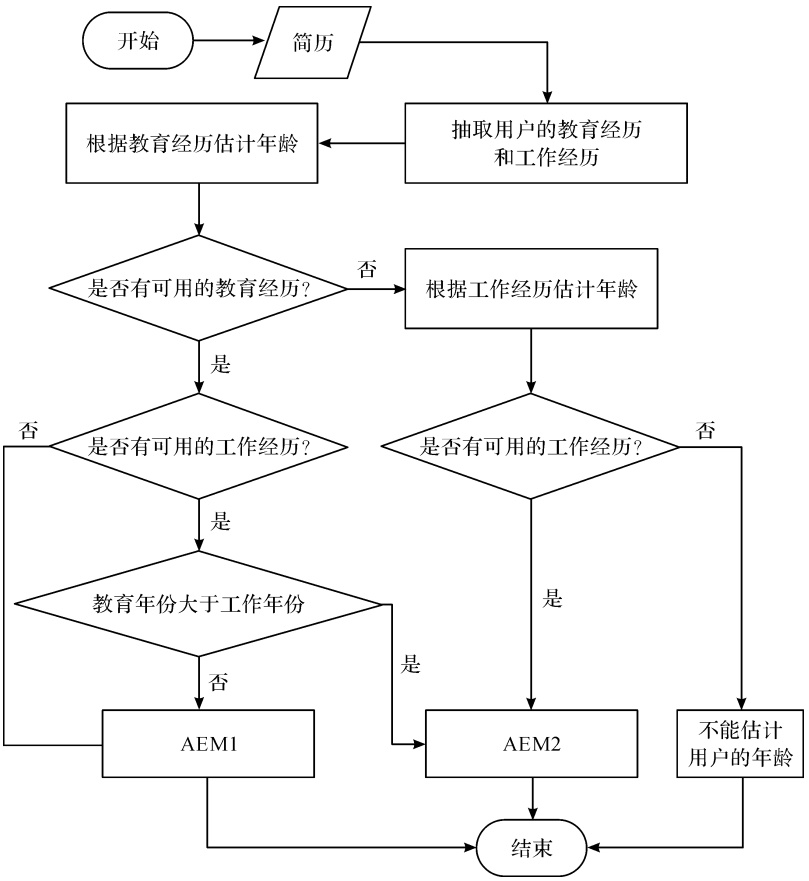

图 1　年龄估计过程

Fig. 1　Age estimation process

3　试验验证

本研究设计了 2 个试验来验证 AEM 的有效性和准确性。

3.1　试验 1

我们已经在数据源中标记了 20 000 个用户的年龄，在数据源中除人工标注年龄的数据之外随机选取 100 个用户，用 AEM 估计他们的年龄。同时仍采用 5 人（A、B、C、D、E）小组人工标注用户年龄，最后比较分析两者的差。试验的部分结果见表 2，其中 y_T 为人工标注年龄的均值（视作“真实年龄”），y 为 AEM 的估计值，d_1 为两者差值的绝对值。

表 2　估计年龄和“真实年龄”的比较

Table 2　Comparison of real ages and estimated ages　　岁

用户 ID	A	B	C	D	E	y_T	y	d_1
17562612	45	46	46	47	45	45.8	45.0	0.8
45517243	45	44	47	46	45	45.4	44.6	0.8
130925345	48	47	47	47	49	47.6	50.9	3.3
16123146	61	60	61	61	60	60.6	61.0	0.4
50275252	62	60	60	62	61	61.0	60.2	0.8
46363372	59	57	59	59	59	58.6	59.0	0.4
20508575	65	62	65	66	64	64.4	65.0	0.6
4755066	47	45	47	46	45	46.0	47.0	1.0
1490046	41	40	43	42	42	41.6	41.0	0.6
21816395	56	53	56	55	57	55.4	56.0	0.6
69861612	44	41	39	44	43	42.2	43.6	1.4
1213625	53	53	54	53	55	53.6	53.0	0.6

表 2(续)

用户 ID	A	B	C	D	E	y_T	y	d_1
53872461	60	57	59	59	59	58.8	60.0	1.2
14016121	38	34	36	35	36	36.0	33.6	2.4
65744103	32	33	33	33	37	33.6	32.0	1.6
165439021	33	32	35	32	32	33.4	33.0	0.4
7565549	38	37	36	38	37	37.2	38.0	0.8
41802332	68	65	63	66	65	65.4	68.0	2.6
74680488	35	35	37	36	37	36.0	35.0	1.0
685144	47	47	47	47	46	46.8	47.0	0.2
173316	44	43	43	43	43	43.2	44.0	0.8
68673379	31	31	33	31	31	31.4	31.0	0.4
95752531	34	34	33	30	36	33.4	38.1	4.7
62634346	47	47	46	48	47	47.0	47.0	0.0
77115020	38	37	37	38	38	37.6	38.0	0.4

AEM 偏差分布见表 3。由表 3 可知,偏差在 3 年以内的占 92%,总体平均偏差为 1.2 岁。AEM 的估计结果与人工分析相近,可见,模型有较高的准确性。

表 3 AEM 偏差分布

Table 3 Distribution of AEM deviation

偏差/岁	百分比/%
[0,1]	87
(1,3]	5
(3,5]	3
(5,∞)	5

3.2 试验 2

人脸识别是近年来年龄估计研究中的常用方法。首先获取试验 1 中随机选取的 100 名用户的简历照片,我们采用微软公司支持的人脸识别技术(http://www.how-old.net/)来估计用户年龄。试验的部分结果见表 4,其中 y_F 为人脸识别年龄估计的值,d_2 为人脸识别估计的偏差。试验结果显示人脸识别估计年龄的平均偏差为 15.6 岁。

表 4 两种年龄估计方法的比较

Table 4 Comparison between two methods of age estimation 岁

用户 ID	y_T	y_F	d_2	用户 ID	y_T	y_F	d_2
17562612	45.8	0	45.8	14016121	36.0	24	12.0
45517243	45.4	31	14.4	65744103	33.6	21	12.6
130925345	47.6	0	47.6	165439021	33.4	0	33.4
16123146	60.6	38	22.6	7565549	37.2	43	5.8
50275252	61.0	0	61.0	41802332	65.4	43	22.4
46363372	58.6	51	7.6	74680488	36.0	25	11.0
20508575	64.4	63	1.4	685144	46.8	44	2.8
4755066	46.0	34	12.0	173316	43.2	0	43.2
1490046	41.6	38	3.6	68673379	31.4	0	31.4
21816395	55.4	31	24.4	95752531	33.4	39	5.6
69861612	42.2	42	0.2	62634346	47.0	29	18.0
1213625	53.6	34	19.6	77115020	37.6	30	7.6
53872461	58.8	47	11.8				

从表 4 中我们还可以看到,人脸识别估计用户年龄的部分结果是 0 或者偏差较大,造成这些结果的原因有很多,如用户照片为空或者非人物头像、照片拍摄时间早、照片模糊、照片被处理等,这些会导致人脸识别方法不能准确地得到用户当前的年龄,例如表 4 中 ID 为“17562612”的用户,事实上在 LinkedIn 的照片为默认空白,因此估计值为 0。

表5 AEM 与人脸识别方法偏差比较

Table 5 Comparison of deviation between AEM and face recognition method

偏差/岁	百分比/%	
	AEM	人脸识别
[0,1]	87	5
(1,3]	5	13
(3,∞)	8	82

AEM 与人脸识别方法偏差分布比较见表 5。由表 5 可知,对 100 名用户年龄估计的结果中,AEM 估计结果偏差在 3 年以内(含 3 年)占比为 92%,而人脸识别方法估

计偏差在此区间占比仅为 18%。同时看到人脸识别方法估计结果偏差大于 3 年占比 82%，远远大于 AEM 的 8%。综合以上分析，AEM 在社交网络用户年龄估计方面，准确性较人脸识别方法更高。

4 结　语

本研究提出了一种年龄估计模型(AEM)，可以根据用户的教育和工作经验来估计其年龄。为了精确估计年龄，我们提出了一个用于数据选择顺序的优先规则，并设计了 2 个试验来评估方法的准确性。试验表明，与人脸识别估计的年龄相比，AEM 具有较高的准确性，也体现了一定的研究价值。用户年龄结合其他相关的数据，能挖掘分析出更有价值的信息。在未来的研究中，我们将致力于实现一个职业推荐系统，可以为用户推荐更有发展潜力、更适合的职业。

参考文献：

[1] HU Z Z, WEN Y G , WANG J F, et al. Facial age estimation with age difference[J]. IEEE transactions on image processing,2017,26(7):3087.

[2] GHAHREMANI P, NIDADAVOLU P S, CHEN N, et al. End-to-end deep neural network age estimation[C]// Interspeech. Hyderabad: ISCA,2018:277.

[3] CHEN S X, ZHANG C J, DONG M, et al. Using ranking-CNN for age estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE,2017:742.

[4] ROTHE R, TIMOFTE R, VAN G L. Deep expectation of real and apparent age from a single image without facial landmarks[J]. International Journal of Computer Vision,2018,126(2/3/4):144.

[5] HUO Z W, YANG X, XING C, et al. Deep age distribution learning for apparent age estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Las Vegas: IEEE,2016:17.

[6] 杨国亮，张雨. 基于卷积神经网络的人脸年龄估计方法[J]. 北京联合大学学报，2018，32(1)：59.

[7] 吴仰波. 基于人脸图像特征表达的年龄估计模型算法研究与实现[D]. 广州：中山大学，2014.

[8] 王华君，惠晶. 基于 CNN 和 LSSVM 的人脸图像年龄估计方法[J]. 信息与电脑(理论版)，2017(7)：85.

[9] 孙劲光，荣文钊. 基于区域的年龄估计模型研究[J]. 计算机科学，2018，45(8)：10.

[10] XING J L, LI K, HU W M, et al. Diagnosing deep learning models for high accuracy age estimation from a single image[J]. Pattern Recognition,2017(66):106.

[11] 胡春龙. 基于人脸图像的头部姿态估计与年龄估计方法研究[D]. 武汉：华中科技大学，2014.

[12] 张珂，高策，郭丽茹，等. 非受限条件下多级残差网络人脸图像年龄估计[J]. 计算机辅助设计与图形学学报，2018，30(2)：346.

[13] 曾雪强，罗明珠，陈素芬，等. 基于自适应多重多元回归的人脸年龄估计[J]. 江西师范大学学报(自然科学版)，2019，43(1)：68.

[14] MCCOMB K, SHANNON G, SAYIALEL K N, et al. Elephants can determine ethnicity, gender, and age from acoustic cues in human voices[J]. Proceedings of the National Academy of Sciences,2014,111(14):5433.

[15] DOBRY G, HECHT R M, AVIGAL M, et al. Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal[J]. IEEE Transactions on Audio Speech and Language Processing,2011,19(7): 1975.

[16] ISELI M, SHUE Y L, ALWAN A. Age-and gender-dependent analysis of voice source characteristics[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Toulouse: IEEE,2006:I-389.

[17] METZE F, AJMERA J, ENGLERT R, et al. Comparison of four approaches to age and gender recognition for telephone applications[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Honolulu: IEEE,2007:IV-1089.

[18] 陈敬. 社交网络中用户年龄识别方法研究[D]. 苏州：苏州大学，2017.

[19] 李源昊，陆平，吴一凡，等. 面向移动社会网络的用户年龄与性别特征识别[J]. 计算机应用，2016，36(2)：364.