UGC 网站用户画像研究①

陈志明, 胡震云

(河海大学 商学院, 南京 211100)

摘 要: 近几年, 社交网络的高速发展使人们的工作、生活、学习方式发生了重大改变, 人们获取知识的方式呈现明显的网络化趋势. 人们通过网络获取信息的同时, 也在其上留下了个人的痕迹, 考虑到现实中获取个人信息成本高昂, 捕捉其在网络中留下的痕迹, 研究其在网络社会中的“映射”, 不失为一种可行的方法. 用户画像作为真实用户的虚拟代表, 是建立在一系列真实数据之上的用户模型. 通过对“知乎”网站的深入挖掘, 构建了基于用户基本属性、社交属性、兴趣属性和能力属性四个维度的动态用户画像模型, 并对“知乎”网站 PM 2.5 话题下 1303 位用户进行实证分析.

关键词: 知乎网; 用户画像; 社交影响力; h 指数

User Portrait Study on UGC Website

CHEN Zhi-Ming, HU Zhen-Yun

(Business School, Hohai University, Nanjing 211100, China)

Abstract: In recent years, the rapid development of social networks has significantly changed the styles of people's work and life. The way people acquiring knowledge shows a clear trend of the network. When people acquire information through the Internet, it also leaves personal traces. Taking the high cost of obtaining personal information in reality into account, it's a good idea to catch people's leaving traces in the network and study its "mapping" in the network society. User portrait as a virtual representative of real users is based on a series of real data on the user model. Based on dynamic modeling of user portrait, the user is defined as the basic attributes, social attributes, interests, and abilities by digging the "ZhiHu" site. Then we make an empirical analysis of the 1303 users who follow the topic of PM 2.5.

Key words: "ZhiHu" site; user portrait; social impact; h-index

近十年, 随着 Web2.0 概念的成熟应用, 社交网络获得了“井喷式”发展, 影响着人们的学习、生活、工作等方式, 每一个“触网”的人都在发生着潜移默化的改变. 人们纵情遨游网络的同时, 也在网络中留下的大量的个人“痕迹”. 随着社交网络规模的不断扩大, 个人的“痕迹”也在不断增多. 在学术界与产业界, 如何获取这些“痕迹”, 如何利用这些“痕迹”的讨论不绝于耳. 社会学家需要这些“痕迹”来剖析网络结构的演变、网络传播学等; 网站的拥有者希望利用“痕迹”为用户提供更好的网络体验; 社交网络上的商家希望利用“痕迹”进行精准的广告投放; 政府部门需要对社交网络上的用户言论进行监管, 尤其是对舆情的控制和非法言论的处理. 用户画像模型为解决这些问题提供了可能的方法.

随着技术的发展和数据资源的累积, 碎片化的“痕迹”才慢慢组合为用户画像. 用户画像作为真实用户的虚拟代表, 是建立在一系列真实数据之上的用户模型. 目前有许多关于用户画像的实际应用, 比如赵曙光[1]基于对高转化率的社交媒体用户研究重要性的认识, 通过一对一的深度访谈的方式对用户的社交媒体使用动机和行为进行总结概括, 对社交用户进行分类. 刘海[2]等基于 4C 理论构建了“用户画像”数据库, 通过对数据库的挖掘来进行消费者群体细分. 在此基础上, 从营销的角度构建了精准营销细分模型. 此外,

① 收稿时间:2016-04-21;收到修改稿时间:2016-05-26 [doi:10.15888/j.cnki.csa.005543]

在个性化推荐系统[3]、产品研发[4]、广告投放[5]等领域也有用户画像的应用. 对用户画像的分析不仅可以更好的认识网络中的用户, 从而低成本、高效率的完成公众参与社会活动, 还可以对现有网络进行更好的完善, 推动社交网络的进一步发展. 因此, 用户画像的构建, 在学术界与产业界都具有重大意义.

1 用户画像模型

用户画像又称用户角色, 在模型的构建过程中往往会以最为浅显和贴近生活的话语将用户的属性、行为和偏好联结起来, 作为实际用户的虚拟代表, 用户画像所形成的角色模型并不能脱离实际场景之外被构建出来. 一个用户可以从多个方面去刻画, 即用户模型可以从多个维度去考虑. “知乎”作为社交化问答网站, 用户在平台上通过某些行为(如回答问题、点击图片、浏览信息流、关注等)生产或获取信息, 也通过其它一些行为(如转发、点赞、评论等)将信息传播出去. 由此, 我们依据社交网络的特性, 结合“知乎”网用户的特点, 将用户画像模型进行四个维度的划分, 即用户的自然属性、社交属性、兴趣属性和能力属性. 同时, 用户在网络社会中的“痕迹”越多, 用户画像模型越能准确反映现实社会中该用户的特征. 但是, 考虑到成本及隐私, 构建“完整”的用户模型几乎不可能, 要结合实际的应用场景, 构建满足条件的用户画像模型即可.

1.1 自然属性

自然属性是指人的肉体存在及其特性, 是人存在的基础. 自然属性指相对稳定和静态的人口属性, 例如: 性别、地域、受教育程度、职业经历等, 由于用户对个人隐私的有意保护, 这些信息的收集主要依靠网站自身的引导、调查、第三方提供等, 并在此基础上进行补充和交叉验证.

以“知乎”为例的自然属性指标如表 1 所示.

表 1 自然属性指标

自然属性指标	简介
用户名	所在网站的识别码
性别	通常以♂表示“男”, ♀表示“女”
个人简介	用一句话介绍自己
居住地	目前所在城市
行业/职业经历	涉及的行业、个人的职业
教育经历	高等教育院校经历或者目前在读院校

用户的自然属性指标在不同的应用场景下对用户画像的描述具有一定程度影响力, 但出于隐私保护的考虑, 往往获取成本较高, 多用于对样本整体进行评价. 其中, 如性别、居住地、行业和简介等指标不具备等级差别, 如果用户的以上指标与用户画像的应用场景相关, 则定义一个函数 $Nolev(i)$ 表示用户 i 的这些属性对构建用户画像的影响:

$$Nolev(i)=\Psi(Gen(i),Pla(i),Tra(i),Abs(i)) \tag{1}$$

其中, $Gen(i)$ 表示用户 i 的性别属性在一定的应用场景下的影响; $Pla(i)$ 表示 i 的居住地属性带来的影响; $Tra(i)$ 表示 i 的行业属性带来的影响; $Abs(i)$ 表示用户 i 的简介对构建用户画像的影响. 这些属性在一定的应用场景下, 只有相关和不相关两种状态, 我们以居住地属性为例定义其影响函数:

$$Pla(i)=\begin{cases}0 & \text{用户与场景无关}\\ pla(i) & \text{用户与场景相关}\end{cases} \tag{2}$$

其中, $pla(i)$ 表示居住地属性的影响值, 其与实际应用场景有关. 同理, 依据式(2)我们可得性别、行业和简介的影响函数:

$$Gen(i)=\begin{cases}0 & \text{用户与场景无关}\\ gen(i) & \text{用户与场景相关}\end{cases}$$

$$Tra(i)=\begin{cases}0 & \text{用户与场景无关}\\ tra(i) & \text{用户与场景相关}\end{cases}$$

$$Abs(i)=\begin{cases}0 & \text{用户与场景无关}\\ abs(i) & \text{用户与场景相关}\end{cases}$$

对于用户自然属性中的教育经历, 不仅影响着用户画像的构建, 还具有等级之分. 本节选用三角模糊数两级比例法对定性指标进行量化.

虽然由于各种原因, 获取完整的用户的自然属性信息困难重重, 但用户的自然属性反映着用户的基本情况, 是连接线上社交网络和线下真实社会的纽带, 其重要性不言而喻. 通过以上几个方面的分析, 定义函数 $Nat(i)$ 表示用户的自然属性, 则:

$$Nat(i)=\Psi(Nolev(i),Edu(i)) \tag{3}$$

1.2 社交属性

本文所探讨的用户的社交属性, 主要通过用户的社交影响力进行衡量, 即用户由于和其他人或团体之间的交互而改变自身观点、情感、态度和行为的现象[6]. 本节基于社交影响力的定义, 综合考虑网络拓扑结构及社交节点的相互影响程度, 对社交影响力进行如下因素分析:

1) 用户的活跃度. 反映了用户的活跃程度, 用户

越活跃, 影响其他用户的机会就越大. 包括用户关注的人数、关注的话题数、关注的专栏数、提问及回答的数量等. 这些都是UGC网站中用户活跃度的直接体现. 考虑到这五个指标有可能不在一个数量级上, 给用户活跃程度的对比带来困难, 因此, 需将它们进行归一化.

$$a_i = \frac{a_i^{'}}{a_{max}} \tag{4}$$

其中, $a_i^{'}$ 表示用户 i 关注的人数; a_{max} 表示用户中关注的人数的最大值. 由公式(4)同理可得:

$$b_i = \frac{b_i^{'}}{b_{max}};\ c_i = \frac{c_i^{'}}{c_{max}};\ d_i = \frac{d_i^{'}}{d_{max}};\ e_i = \frac{e_i^{'}}{e_{max}};$$

其中, a_i, b_i, c_i, d_i, e_i 分别表示归一化后的用户关注的人数, 用户关注的话题数, 用户关注的专栏数, 用户的提问数, 用户的回答数. 则:

$$\omega(i) = \alpha_1 a + \alpha_2 b + \alpha_3 c + \alpha_4 d + \alpha_5 e \tag{5}$$

其中, $\omega(i)$: 用户 i 的活跃度的影响力权重; $\alpha_i (i = 1,2,3,4,5)$: 权重系数, 设定在 $[0,1]$ 范围, 且 $\sum \alpha_i = 1$.

2) 用户“邻居”的影响力. 定义“邻居”为用户的关注者, 等同于“粉丝”. 社交网络中信息的流动离不开“邻居”, 邻居节点作为传播的载体, 本身的影响力同样重要. 对于“邻居”的影响力, 我们使用“邻居”的关注者数量及“邻居”与用户的亲密度进行度量.

社交网络的拓扑结构用图 $G = (V, E)$ 表示. 其中 $n = |E|$ 表示节点数; e_{ij} 表示节点 i 和 j 之间的边; $A_{n \times n}$ 表示图的连接矩阵; a_{ij} 是其中的元素; ω_{ij} 表示节点 i 和 j 之间的权重.

定义节点 i 的关注者集合为:

$$N(i) = \{ j \mid (i,j) \in E \}$$

则邻居节点 j 对 i 的影响力为:

$$Inf_1(i,j) = G(num(i,j), int(i,j)) \tag{6}$$

其中: $num(i,j)$ 表示 j 在关注者数量方面对 i 的影响程度, 令 $|T(j)|$ 为 j 的关注者总数, $\sum_{a \in N(i)} |T(a)|$ 为 i 的所有邻居节点的关注者总数, 则:

$$num(i,j) = \frac{|T(j)|}{\sum |T(a)|} \tag{7}$$

$int(i,j)$ 表示 j 在亲密度方面对 i 的影响程度, 令 $F(i,j)$ 表示用户 i 与其邻居 j 相关话题下的交互集合. 则:

$$int(i,j) = \frac{|F(i,j)|}{\sum |F(i,j)|} \tag{8}$$

其中, $|F(i,j)|$ 表示集合中元素个数.

借鉴 PageRank 的核心思想, 本节关于用户“邻居”影响力的定义如下:

$$\omega(i,j)^{k+1} = d + (1-d) \sum_{j \in N(i)} [\omega(i,j)^k \times V_{ij}] \tag{9}$$

其中: $\omega(i,j)^{k+1}$ 与 $\omega(i,j)^k$ 表示更新前后的用户“邻居”影响力, d 是阻尼系数, 与算法的收敛速度有关, 经验值取 0.15. V_{ij} 为投票矩阵, 表示节点的影响力权重, 这里的权重指的是节点在关注者数量方面的权重和节点在亲密度方面的权重. V_{ij} 中元素 v_{ij} 按如下公式计算, 即投票权重为投票用户在关注者数量方面的影响力权重乘以其与被投票用户的亲密度方面的影响力权重占该用户所有出边权重的比重:

$$v_{ij} = \begin{cases} 0, & (i,j) \notin E \\ \dfrac{num(i,j) \times int(i,j)}{\sum_{(k,j) \hat{I} E} int(k,j)}, & (i,j) \in E \end{cases} \tag{10}$$

综合考虑用户本身的活跃度与用户“邻居”的影响力, 且这两者均与社交影响力成正比, 则可得用户 i 在网络中的全局社交影响力, 即用户的社交属性为:

$$Inf(i) = \omega(i) \times \omega(i,j)^{k+1} \tag{11}$$

1.3 兴趣属性

网站会在用户注册时要求其选择自己感兴趣的话题, 并依此进行相关内容的推荐, 因此用户所关注的话题可以看作是其显性兴趣; 用户在浏览网站的过程中, 会对自己感兴趣的话题进行提问、回答与收藏, 因此用户的这些行为也能反映其兴趣, 称之为隐性兴趣. 下面我们对兴趣属性进行建模:

1) 显性兴趣建模

用户会对自己感兴趣的话题进行关注, 以方便获取这方面的相关内容, 所以我们可以将用户的关注话题看作显性兴趣的反映. 对于兴趣标签的权重计算, 我们使用 TF-IDF 方法, TF-IDF 是一种经典的信息加权技术, 其值常用来度量一个词语在文件中的地位[7]. 通过用户的话题标签表示用户的兴趣模型, 标签映射的回答数即为标签被使用的次数, 利用 TF-IDF 方法计算用户 i 的兴趣标签 t_j 的权重 ω_j:

$$\omega_j = \frac{tf_{ij}}{\sum_{t_k \in D_i} tf_{ik}} \times \log\left(\frac{N}{n_j}\right) \tag{12}$$

其中, tf_{ij} 表示用户 i 使用标签 t_j 的次数, N 为用户集

合中用户的总数，n_j 为用户集合中关注标签 t_j 的用户数，D_i 表示用户 i 的兴趣标签集合. 对其进行归一化处理，即:

$$\omega_j^x = \frac{\omega_j}{\omega_{max}}$$

其中，ω_j^x 为显性兴趣标签 t_j 的权重，ω_{max} 为显性兴趣标签的最大值. 则用户 i 的显性兴趣模型表示为:

$$\left\{(t_1,\omega_1^x),(t_2,\omega_2^x),\text{L},(t_n,\omega_n^x)\right\} \quad (13)$$

2) 隐性兴趣建模

用户的关注话题，直观体现了用户的显性兴趣. 然而话题标签是用户注册时人为设定的，很多用户为了节省注册时间，会任意勾选话题，又或者选择很多话题，这就造成了用户兴趣度量的准确性不高. 隐性兴趣不同于显性兴趣，兴趣标签无法根据用户的关注话题直接获得，而是通过用户的交互行为获取. 我们通过分析用户的提问、回答和收藏来构建用户的隐性兴趣. 关于它们的定义如表 2 所示.

表 2 用户行为定义

行为名称	行为函数	兴趣标签集合	行为权重
提问	$Question(i)$	T_1	λ_1
回答	$Answer(i)$	T_2	λ_2
收藏	$Save(i)$	T_3	λ_3

则用户的隐性兴趣标签权重为:

$$\omega_j^y = \Gamma(Question(i), Answer(i), Save(i)) \quad (14)$$

其中，$Question(i)$ 为用户 i 进行提问时的行为函数，其产生的兴趣标签集合为 T_1，定义提问行为本身的权重为 λ_1，隐射到标签集合中即来自于 T_1 的标签权重为 λ_1；同理，$Answer(i)$ 为用户 i 进行回答时的行为函数，其产生的兴趣标签集合为 T_2，来自于 T_2 的标签权重为 λ_2；$Save(i)$ 为用户 i 进行收藏时的行为函数，其产生的兴趣标签集合为 T_3，来自于 T_3 的标签权重为 λ_3. 且定义 $\lambda_1 = \lambda_2$，$\lambda_1 + \lambda_2 + \lambda_3 = 1$，由于收藏更能反映用户偏好，定义 λ_3 f 0.5. 下面我们对用户隐性兴趣建模步骤进行详细表述:

① 对用户 i 的提问进行爬虫抓取，获得提问标签集合 T_1，$N_1 = |T_1|$ 表示集合中所包含的标签个数. 对用户 i 的回答与收藏内容进行相同的操作，$N_2 = |T_2|$ 表示回答内容中所包含的兴趣标签个数，$N_3 = |T_3|$ 表示收藏内容中的标签个数. 定义 $N_i = N_1 + N_2 + N_3$ 为用户 i 的行为中的全部兴趣标签个数.

② 对用户 i 的提问标签集合进行整理，将其中的相同标签聚在一起，设 $t_{1j} \in T_1$ 表示集合 T_1 中的标签 t_j，n_{1j} 表示集合 T_1 中的标签 t_j 的个数，则对于集合 T_1 中的标签 t_j 的兴趣权重为:

$$\omega_{1j} = \lambda_1 \times \frac{n_{1j}}{N_i}$$

③ 对用户 i 的回答标签进行聚类，设 $t_{2j} \in T_2$ 表示集合 T_2 中的标签 t_j，n_{2j} 表示集合 T_2 中的标签 t_j 的个数，则对于集合 T_2 中的标签 t_j 的兴趣权重 ω_{2j} 为:

$$\omega_{2j} = \lambda_1 \times \frac{n_{2j}}{N_i}$$

同理：$\omega_{3j} = \lambda_1 \times \frac{n_{3j}}{N_i}$. 由以上步骤可得到用户 i 的标签 t_j 的兴趣权重 ω_j 为:

$$\omega_j = \omega_{1j} + \omega_{2j} + \omega_{3j}$$

④ 标签 t_j 的兴趣权重 ω_j 表示的是一种词频权重，借鉴上节的 TF-IDF 算法，我们对 ω_j 进行修正:

$$\omega_j^{'} = \omega_j \times \log(\frac{N}{n_j})$$

其中，$\omega_j^{'}$ 是对 ω_j 修正后的权重，N 为用户集合中用户的总数，n_j 为用户集合中关注标签 t_j 的用户数.

⑤ 对标签 t_j 的兴趣权重 $\omega_j^{'}$ 进行处理:

$$\omega_j^y = \frac{\omega_j^{'}}{\omega_{max}^{'}}$$

其中，ω_j^y 表示用户的隐性兴趣权重，$\omega_{max}^{'}$ 表示用户 i 的兴趣权重的最大值.

用户的隐性兴趣权重通过提问、回答和收藏来度量，通过对三个行为所涉及的内容进行标签提取，用户 i 的隐性兴趣模型表示为:

$$\left\{(t_1,\omega_1^y),(t_2,\omega_2^y),\text{L},(t_n,\omega_n^y)\right\} \quad (15)$$

3) 用户兴趣建模

本节结合用户的显性兴趣和隐性兴趣构建用户 i 的兴趣模型，在用户兴趣模型中，对于某位用户，其兴趣标签项为 t_j，计算其在模型中的权重 ω_j:

$$\omega_j = \gamma \times \omega_j^x + (1-\gamma)\omega_j^y \quad (16)$$

其中，ω_j^x 为显性兴趣模型中兴趣项 t_j 的权重；ω_j^y 为隐性兴趣模型中的兴趣项 t_j 的权重. 一般情况下，用户的隐性兴趣更能体现用户的兴趣偏好，故令 λ p 0.5. 在得到用户 i 的兴趣项排名后，根据实际情况，取前 m 项兴趣进行研究分析，则用户的兴趣属性模型 $Int(i)$ 表示为:

$$Int(i) = \{(t_1,\omega_1),(t_2,\omega_2),\text{L},(t_m,\omega_m)\} \quad (17)$$

1.4 能力属性

本文中的能力属性特指用户输出优质内容的水平. 用户生产内容(user generated content, UGC)是在Web2.0 环境下出现的一种新兴的网络信息资源创作与组织模式, 泛指以任何形式在网络上存在的由用户发表的文字、图片、视频等内容, 也就是说, 用户既是网络内容的浏览者, 也是网络内容的生产者与传播者[8]. “知乎”作为典型的UGC 网站, 用户优质内容的产出能力极其重要, 它是网站的核心竞争力. 本节综合考虑内容的质与量, 借鉴 Hirsch 教授设计的科学计量评价指标 h 指数(highly cited index)[9]与金碧辉等人提出的R 指数[10], 进行用户能力属性的度量. 其中, Hirsch 将h 指数定义为: 一位作者的h 指数等于其发表了 h 篇至少被引 h 次的论文, 即一个作者的 h 指数表明其至多有 h 篇论文被引用了至少 h 次.

H 指数具备简洁新颖的特点, 但也有自己的不足. 首先, h 指数对高被引论文的影响力反映不足; 其次, h 指数对普通研究者缺乏区分度, 对于大量普通研究者来说, 他们拥有相同的h 指数, 且h 指数的值都比较低; 最后, h 指数受自引和合作发文的影响, 大量自引可以显著改变h 指数. 针对h 指数的缺陷, 金碧辉提出了 R 指数. R 指数表示的是h 核内论文的总被引频次的平方根. R 指数的数学公式如下:

$$R=\sqrt{\sum_{j=1}^{h} cit_j}$$

式中 cit_j 表示绩效核内第 j 篇论文的被引频次. cit_j 至少等于 h 值. R 指数在不改变 h 核形态的前提下, 对 h 核内部论文的被引频次进行计算, 其度量结果可以有效区分同值 h 指数.

本节通过分析 h 指数与 R 指数各自的特点后, 将两种指数配对使用, 将会有效弥补 h 指数的不足, 对用户的能力属性进行度量如下所示:

1) 回答能力指数(A_I). 表示用户回答问题的数目与其获取的赞同及讨论之间的关系. 根据 h 指数的定义, 对用户回答的 n 个问题组合而言, 其中有 h 个回答每篇至少获得赞同数 h 次, 剩下的 $(n-h)$ 个答案每个被赞同的次数都少于 h 次, 则 h 为该用户的 h 指数, h 个答案的被赞同次数总和的平方根为该用户的 R 指数. 设 r 是赞同次数降序排列的答案的序次, ZT_r 是回答 r 获得的赞同数, 则有以下序列:

$$\begin{aligned} &r=(1,2,\cdots,r,\cdots z)\\ &ZT=(ZT_1,ZT_2,\cdots,ZT_r,\cdots,ZT_z)\\ &ZT_1\ge ZT_2\ge\cdots\ge ZT_r\ge\cdots\ge ZT_z \end{aligned} \tag{18}$$

赞同的 h 指数和 R 指数理论上就是:

$$h_Z=max\{r:r\le ZT\} \tag{19}$$

$$R_Z=\sqrt{\sum_{j=1}^{h_Z} ZT_j} \tag{20}$$

定义向量 $\vec{Z}=(h_Z,R_Z)$ 为赞同的指数向量.

同理: 讨论的 h 指数和 R 指数理论上就是:

$$h_t=max\{r_t:r_t\le TL\}$$

$$R_t=\sqrt{\sum_{j=1}^{h_t} TL_j}$$

定义向量 $\vec{T}=(h_t,R_t)$ 为讨论的指数向量. 其中: r_t 表示讨论次数降序排列的答案的序次, TL_j 表示回答 r 获得的讨论总数.

综上: 用户 i 回答能力指数向量为:

$$\overrightarrow{A_I(i)}=\alpha\vec{Z}+(1-\alpha)\vec{T}=(\alpha h_z+(1-\alpha)h_t,\alpha R_z+(1-\alpha)R_t) \tag{21}$$

其中, α 表示赞同指数在回答能力指数所占的权重.

2) 提问能力指数(Q_I). 表示用户提问的数目与其获取的答案数及关注人数之间的关系. UGC 网站中, 提问者与回答者同等重要, 他们都是网站内容的产出者, 存在着一种辩证统一的关系. 根据上文h 指数及R 指数的定义, 设 m_d 表示答案数目降序排列的提问的序次, DA_m 表示提问 m 获得的答案总数, 则答案的 h 指数与 R 指数为:

$$h_d=max\{m_d:m_d\le DA\}$$

$$R_d=\sqrt{\sum_{j=1}^{h_d} DA_j}$$

同理: 关注人数的 h 指数与 R 指数为:

$$h_g=max\{m_g:m_g\le GR\}$$

$$R_g=\sqrt{\sum_{j=1}^{h_g} GR_j}$$

由上文分析可知: 用户 i 提问能力指数为:

$$\overrightarrow{Q_I(i)}=\beta\vec{D}+(1-\beta)\vec{G}=(\beta h_d+(1-\beta)h_g,\beta R_d+(1-\beta)R_g) \tag{22}$$

其中, $\vec{D}$ 表示答案的指数向量, $\vec{G}$ 表示关注的指数向量, β 表示答案数目在提问能力指数中所占权重.

提问能力与回答能力在UGC网站中同等重要, 故本节关于用户 i 的能力属性定义如下:

$$Abi(i)=\overrightarrow{A_I(i)}+\overrightarrow{Q_I(i)} \tag{23}$$

其中, $Abi(i)$ 表示用户的能力属性, 需要注意的是, R 指数的意义是为了区分相同h 指数下能力属性的大小, 即 R 的值只是在 h 指数的值相同的情况下起作用.

1.5 用户画像模型构建

以上四节分别从用户的自然属性、社交属性、兴

趣属性及能力属性四个方面对用户画像模型进行构建, 该模型可以较为全面的对用户进行模型抽象, 但是众所周知, 用户画像模型的应用离不开实际的应用场景, 在面对不同的场景时, 用户画像所侧重的属性是不同的. 这就要求模型具有动态特征, 定义函数 $Per(i)$ 表示用户的画像模型, 则:

$$Per(i)=f(\theta_1,\theta_2,\theta_3,\theta_4)f(Nat(i),Inf(i),Int(i),Abi(i)) \quad (24)$$

其中, $Nat(i),Inf(i),Int(i),Abi(i)$ 分别度量用户 i 的自然属性, 社交属性, 兴趣属性及能力属性; $\theta_i,i=1,2,3,4$ 表示属性在实际应用场景中的权重, 且 $\theta_1+\theta_2+\theta_3+\theta_4=1$.

2 用户画像实证分析

众所周知, 用户画像的应用离不开使用场景的设置, 没有使用场景, 空谈用户画像是没有实际意义的. 我们对用户画像模型的构建过程有了深刻的理解, 接下来我们将结合实际的场景设置, 对用户画像模型的实际应用进行研究. 本文所用数据集为“知乎”网站关注 PM2.5 话题的用户的数据信息. 截止到 2015 年 12 月, 共有 1318 人关注该子话题, 数据由 Gooseeker 爬虫抓取, 其中成功抓取 1303 位用户数据, 成功率为 98.9%. 本文的实证即对这 1303 位用户进行分析.

场景一: 网站核心用户甄别

对于 UGC 网站而言, 其核心用户应具备输出优质内容的能力, 即用户的能力属性值排名靠前. 由 1.4 节可知, 用户的能力属性包括用户的提问能力及回答能力, 分别通过提问能力指数和回答能力指数进行度量. 令 α=0.4, β=0.6, 可得用户能力属性的散点图如图 1 所示.

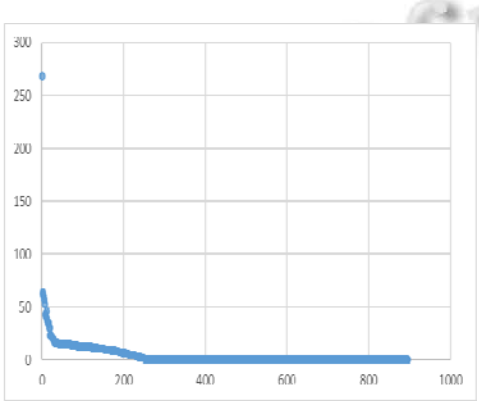

图 1 用户能力属性散点图

分析上图可知, 数据集中绝大部分用户的能力属性值比较小, 主要原因是其提问与回答数都比较小, 或者其少量的提问与回答中, 质量并不高, 所以并没有得到网络中用户的认同; 在图中, 有一位用户的能力属性 h 值高达 268.4, 该用户在网络中的回答数量为 1417, 提问数量为 106, 而其得到的总赞数为 653528; 同样, 用户 AreYouKiddingMe 的 h 值为 61.8, 可其回答数量为 2031, 提问数量为 505; 可见 h 值可以更好的反映用户的能力, 它不仅考虑了用户输出内容的量, 还考虑了内容的质.

场景二: 公众参与环保政策的制定

公众参与环保政策的制定, 需要考虑两个方面的问题, 一方面是公众的选择问题, 另一方面是公众的高效参与问题.

关于公众的选择, 可以应用用户画像模型得到结果. 首先, 评估用户的教育经历, 选取学历为本科及以上的用户; 其次, 分析用户能力属性, 能力属性值越大, 表示其输出的内容质量越高; 然后, 结合用户的兴趣属性, 判断其输出内容是否与环保相关; 最后即可确定该用户是否适合参与环保政策的制定. 根据以上分析, 式(24)转变为:

$$Per(i)=f(\theta_1,\theta_3,\theta_4)f(Nat(i),Int(i),Abi(i))$$

分析数据集可得到部分结果, 如表 3 所示. 需要注意的是, 表格中的能力值是结合兴趣属性后在环境保护相关话题下的能力, 是对环保相关的提问、回答的度量.

表 3 场景二分析结果

用户名	教育经历	兴趣话题	能力值
土豆冰激凌	河海大学环境工程	环境保护; 空气质量	2.8
心向蓝天	无	环境保护; 环境监测	2.7
PE Vincent	无	空气净化器; 雾霾	2.0
刘虓震	无	教育; 环境保护	1.6
Tian Tony	北卡大学环境科学	环境污染; PM 2.5	1.5
轩净	无	空气净化器; 空气质量	1.4
Joshua Chang	复旦大学环境化学	PM 2.5; 生活	1.2
江宁	中国药科大学环境科学	电影; 环境保护	0.7

由上表可以看出, 教育经历在一定程度上影响着用户的兴趣. 在确定了哪些公众适合参与政策制定的情况下, 需要考虑公众的高效参与问题. 本文以目前我国公民的主要参与方式为出发点进行论述.

1) 关键公众参与法. 即寻找与政策制定相关的公民进行咨询, 上表中选取的关键公众, 有效弥补了传统选择方法中只关注精英分子的缺陷, 真正做到让普

通大众参与到环保相关政策的制定中.

2) 公民调查法. 即运用问卷调查或现场访谈的形式, 了解公众对于相关政策的态度和看法. 在新媒体时代, 利用用户画像模型将网络问卷发放给特定的公众, 既提高了调查的有效性, 又降低了相关工作人员的时间成本.

3) 民众论坛. 即将网络中适合参与环保政策制定的民众组织起来, 构建专业的民众论坛. 首先, 为公众参与提供通道; 其次, 引导舆论走向, 构建官方与民间的沟通渠道; 然后, 搭建专业型平台, 为普通公众的环保问题提供解决方案; 最后, “取之于民”的政策更利于推行.

参考文献

1 赵曙光.高转化率的社交媒体用户画像:基于 500 用户的深访研究.现代传播:中国传媒大学学报,2014,(6):115–120.

2 刘海,卢慧,阮金花,田丙强,胡守忠.基于“用户画像”挖掘的精准营销细分模型研究.丝绸,2015,52(12):37–42.

3 邢星.社交网络个性化推荐方法研究[博士学位论文].大连:大连海事大学,2013.

4 余孟杰.产品研发中用户画像的数据模建——从具象到抽象.设计艺术研究,2014,(6):60–64.

5 Bakshy E, Eckles D, Yan R, Rosenn I. Social influence in social advertising: Evidence from field experiments. Proc. of the 13th ACM Conference on Electronic Commerce. ACM. 2012. 146–161.

6 Rashotte L. Social influence. The blackwell encyclopedia of social psychology, 2007, 9: 562–563.

7 宗成庆.统计自然语言处理.北京:清华大学出版社,2008.

8 朱庆华.新一代互联网环境下用户生成内容的研究与应用.北京:科学出版社,2014.

9 Hirsch JE. An index to quantify an individual’s scientific research output. Proc. of the National academy of Sciences of the United States of America, 2005, 102(46): 16569–16572.

10 金碧辉.R 指数,AR 指数:H 指数功能扩展的补充指标.科学观察,2007,2(3):1–8.