Software Engineering and Applications 软件工程与应用, 2020, 9(3), 201-215
Published Online June 2020 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2020.93024

Hans汉斯

Research on Character Portrait Technology Based on Text Analysis

—Taking the Influencer as an Example

Lu Zhang, Yu Chen, Jiaxin Jing, Jinglun Cai

University of International Relations, Beijing
Email: 294815113@qq.com

Received: May 21st, 2020; accepted: Jun. 3rd, 2020; published: Jun. 10th, 2020

Abstract

Social media high-impact users combine unique content, the commercial value of their own topics, and efficient traffic monetization capabilities. Constructing portraits of high-impact characters can directly display the typical characteristics of group personnel, which plays an important role in expanding the radiating power of excellent network culture, providing accurate services to the platform, maintaining core users, and supervising public opinion. Taking influencer as an example, we use Python to acquire and process user behavior data, construct user portrait conceptual models from multiple dimensions of influencer, use tools such as "WordCloud", "Mapplotlib", and "Pyecharts" to visualize and develop Empirical Research. The experimental results show that influencer is mainly divided into two categories. Taking randomly selected domains as typical representatives, we obtain typical portraits of different high-impact users. According to the key group characteristics of users and their fans, it shows that the user portrait model can bring tremendous application value in the aspects of personalized service, recommendation system and precision marketing of social platforms.

Keywords

High Impact Users, Text Analysis, Figure Portrait, Zhihu

基于文本分析的人物画像技术研究

——以知乎大 V 为例

张　璐，陈　宇，景嘉欣，蔡京伦

国际关系学院，北京
Email: 294815113@qq.com

文章引用: 张璐, 陈宇, 景嘉欣, 蔡京伦. 基于文本分析的人物画像技术研究[J]. 软件工程与应用, 2020, 9(3): 201-215.
DOI: 10.12677/sea.2020.93024

收稿日期：2020年5月21日；录用日期：2020年6月3日；发布日期：2020年6月10日

摘　要

社交媒体高影响力用户集独特的内容、自带话题的商业价值以及高效的流量变现能力于一体。构建高影响力人物画像，可以直观展示群体人员的典型特征，对扩大优秀网络文化的辐射力以及对平台提供精准服务、维系核心用户、监管引导舆情等具有重要作用。我们以知乎大V为例，利用Python获取和处理用户行为数据，从知乎大V多个维度构建用户画像概念模型，用“WordCloud”、“Mapplotlib”以及“Pyecharts”等工具进行可视化，开展实证研究。实验结果表明知乎大V主要分为两大类，以随机选取的领域大V作为典型代表，得到不同高影响力用户的典型人物画像，依据用户及其粉丝的关键群体特征属性，表明用户画像模型可以在社交平台的个性化服务、推荐系统、精准营销等方面带来巨大的应用价值。

关键词

高影响力用户，文本分析，人物画像，知乎

1. 引言

(一) 研究背景

凯度最新向全球发布的《2020 年媒介趋势与预测报告》(2020 Media Trends & Predictions Report)指出，明年整个行业将遇到“数字化悖论”[1]。报告表示，为走出这种困局，营销人需要理解消费者对于更相关、更个人化内容的需求，但与此同时不能破坏它对品牌的信任，也不能侵犯消费者隐私[1]。在此背景下，高影响力用户作为信息传播的关键群体，其观点及言论深刻影响其他用户的思维方式，加速社交网络信息传播、增强用户间有效的交互活动。同时，粉丝经济催生高影响力用户实现其经济价值与社会价值。因此，对高影响力用户构建人物画像，识别其典型特征及潜在需求，对扩大优秀网络文化的辐射力和感染力及平台提供精准服务、维系核心用户、监管引导舆情等至关重要。然而，凯度的研究显示，消费者对于“个性化”概念的反应正在变得两极化[2]。报告认为，个性化定制的信息不应被当作完成销售的捷径，而是长期策略的基础——用来获得消费者的信任，增加对品牌的忠诚度。

(二) 研究对象

本文以网络问答社区知乎的高影响力用户作为研究对象，通过对高影响力用户及其粉丝群体画像，试图了解高影响力用户内部的数据特征，并解决如何有效进行知乎数据挖掘的问题，进而扩展知乎数据挖掘在商业情报，社会研究，舆情研究等领域的应用。知乎是近年来新兴的以分享彼此的专业知识和经验见解为理念，以保持严谨、理性的社区氛围为特征的网络问答社区。知乎是网络问答社区，连接各行

[1] 数字化悖论：一方面层出不穷的新媒体渠道似乎都能带来新的机遇；另一方面数字媒介渠道数量过多，触及消费者的可能性反而下降了。
[2] 两极化：有 45%的人表示向自己精准投放的广告比其他广告更有趣，同时也有超过半数的(54%)人反对根据自己过往的网络行为向自己投放广告。

各业的用户。用户分享着彼此的知识、经验和见解，为中文互联网源源不断地提供多种多样的信息。截至 2018 年 11 月底，用户数破 2.2 亿，其问题数超过 3000 万，回答数超过 1.3 亿[3]。知乎在问答社区基础上融入社交元素，建立了全新的内容创造与传播机制，其高质量的社区内容已逐渐成为互联网用户获取知识的重要途径。从这些指标可以判断，知乎已经是国内在线问答社区的标杆，而其中高影响力用户庞大的粉丝群体，以及高质量的回答更对知乎平台有举足轻重的意义，庞大的信息资源和活跃用户的积累是本文选择知乎为研究对象的原因。

(二) 研究目的及意义

1) 研究目的

本文将用户画像的理念应用到社交媒体高影响力用户群体的分析，将传统的单一高影响力用户追踪转变为对群体的抽象概括，以用户与粉丝双方的信任机制为支撑，人格、内容、流量分别对应用户基本信息、行为、价值属性，融合多维属性构建高影响力人物画像可视化模型。

2) 研究意义

社交媒体高影响力用户集独特的内容能力、魅力的人格化特征、自带话题的势能价值及高效的流量变现能力于一体。构建高影响力人物画像，直观展示群体人员典型特征，对扩大优秀网络文化的辐射力和感染力及平台提供精准服务、维系核心用户、监管引导舆情等具有重要作用。

3) 研究方法

研究思路：收集文本数据，数据分析，按照符合人类情感的方向描摹用户画像。

具体方法：在使用 Scrapy 爬虫收集知乎内部数据库文本数据的基础上，对文本数据进行清洗，再将清洗后的数据进行数据集成，规约，使用 jieba 分词将文本切分为成词词语。然后使用 Python 下的 snowNLP 类库对词库进行情感分析得出正负样本，利用新的数据得出情感分类模型。再采用 Python 自带的工具 pyecharts、matplotlib 帮助我们做数据的可视化处理，探究该用户群体的心理情感特征和用户行为。通过横向对比数据项建立 5 个维度的用户画像模型，从而得出高质量用户画像。

2. 相关概念及基础理论

(一) 大 V

大 V，又称意见领袖或舆论领袖，是指在人际传播网络中经常为他人提供信息，同时对他人施加影响的“活跃分子”，他们在大众传播效果的形成过程中起着重要的中介或过滤的作用，由他们将信息扩散给受众，形成信息传递的两级传播[2]。

新媒体时代下，大 V 成为互联网时代创意的跨界整合，集独特的内容能力、魅力的人格化特征、自带话题的势能价值及高效的流量变现能力于一体，创造平台红利与人格红利。而高影响力用户通过发布、转发及评论等行为展示独特的内容生成能力及魅力的人格化特征，影响其他用户的观点，并使其自发地产生互动行为[3]，以社交营销等方式实现流量变现。

(二) 人物画像

人物画像通常有两种理解。一种叫做 Persona，也叫做用户角色，是描绘抽象一个自然人的属性；一种叫做 Profile，是和数据挖掘、大数据息息相关的应用[4]。而本文所讨论的是后者，通过数据建立描绘用户的标签。随着互联网的发展和信息的快速更新，传统的线下交流模式已无法满足用户多元化、个性化的知识需求和专业性、及时性的服务需求，以知乎、豆瓣、简书等为代表的社交类学术移动应用程序已成为科研用户获取知识资源、进行学术交流的新途径。

[3] 该数据来源于 2018 年 12 月 13 日的知乎官方。截至 2018 年 11 月底，用户数破 2.2 亿，同比增长 102%。据此前数据，2017 年底，知乎注册用户达到 1.2 亿，2018 年 8 月底，知乎用户突破 2 亿。

用户画像(UserProfile)，即用户信息的标签化，是建立在一系列数据之上的目标用户模型。是根据用户基本属性、社会属性、生活习惯和消费行为等信息而抽象出的一个标签化的用户模型[5]。构建用户画像的核心工作即是给用户贴“标签”，而标签是通过对用户信息分析而来的高度精炼的特征标识。用户画像的意义在于了解用户，预测用户的真实需求和潜在需求，精细化地定位人群特征，挖掘潜在的用户群体，为媒体网站、广告主、企业及广告公司提供群体用户的差异化特征。用户画像在精准营销、移动用户行为研究、搜索引擎以及个人信息管理方面都有很多应用。用户画像的应用对社会化问答社区的发展也具有重要的意义[6]。

3. 数据预处理

数据预处理环节有利于提高大数据的一致性、准确性、真实性、可用性、完整性、安全性和价值性等方面质量，而大数据预处理中的相关技术是影响大数据过程质量的关键因素[7]。

(一) 数据获取

本文以知乎为研究样本，使用爬虫 Scrapy 从 PC 端爬取用户数据。随机选取 2020 年 3 月 15 日至 2020 年 3 月 20 日间不同热门类别的高影响力用户作为数据样本，涉及文化、艺术设计、影视三个类别。采集了 3 个高影响力用户的相关信息，合计 12,954 名用户的数据记录。获取了高影响力用户(下文用“大 V”代替)的公开信息，如用户的粉丝数量、关注数量、每日的回答、关注话题、提问的关键字、提问记录数以及收到赞同、收藏或者感谢的数量等。

(二) 数据清洗

数据集涵盖了大 V 的公开信息，数据清洗是构造画像标签的基础。数据清洗与处理单元主要负责将原始和基础数据预处理为后续分析所需的数据，主要包括以下几个内容：

1) 缺失值清洗：确定缺失值范围；去除不需要的字段；填充缺失内容；重新取值。

2) 格式内容清洗：时间、日期、数值、全半角等显示格式不一致；内容中有不该存在的字符；内容与该字段应有内容不符。

3) 逻辑错误清洗：去重(重复的用户 ID)；去除不合理值、修正矛盾内容，例如粉丝个人描述信息不符合实际(年龄失真，个性签名存在歧义，存疑僵尸粉丝等等)。

4) 非需求数据清洗：提取文本时删除无关干扰字眼，例如微博，微信，微信号(与研究主题产生干扰，污染数据)，关注者，无关数字，粉丝等。

(三) 数据处理

将清洗后的数据分词，然后进行数据集成，从而形成集中、统一的数据库，有利于提高大数据的完整性、一致性、安全性和可用性等方面质量。然后对集成后的数据进行归约，即在不损害分析结果准确性的前提下降低数据集规模，运用数据抽样技术对其进行简化，这一过程有利于提高大数据的价值密度，即提高大数据存储的价值性[8]。我们将数据集整理为 16 个数据项，以便于后续构建和分析人物画像。

Table 1. High-impact user data items
表 1. 知乎高影响力用户数据项

编号	CPM 维度	数据名称	注释
a_1	B	用户 ID	用户唯一标识符，为用户属性信息，可直接获取。
a_2	B	研究领域	用户注册时选择的学科分类和研究方向，为用户属性信息，可直接获取。
a_3	P	提问数目	用户提问的数目，与积极程度正相关，可直接获取。
a_4	P	回答问题数目	用户回答的数目，与积极程度正相关，可直接获取。

Continued

a_5	I	用户被赞总数	用户回答被点赞的数目之和，与影响力正相关，可直接获取。
a_6	I	用户粉丝数目	用户的粉丝数量，与影响力正相关，可直接获取。
a_7	I	用户关注数目	用户关注的人数量，与互动程度相关，可直接获取。
a_8	D	用户所关注者的个性签名情感分析(图表)	用户所关注者个性签名分词之后的情感分析可视化图表，与偏好方向相关，通过爬虫获取。
a_9	D	用户所关注者的研究领域(词云)	用户所关注者注册时选择的学科分类和研究方向，与偏好方向相关，通过爬虫获取。
a_{10}	D	用户回答问题的情感分析(图表)	用户所回答问题的题干分词之后的情感分析可视化图表，用于判断用户的情感积极消极程度，与偏好方向相关，通过爬虫获取。
a_{11}	D	用户回答问题(词云)	用户所回答问题的题干分词之后的关键词词云，用于归纳用户回答问题的特征，与偏好方向相关，通过爬虫获取。
a_{12}	F	粉丝 VIP 比例(图表)	用户粉丝的 VIP 占比，属于粉丝群体特征，通过爬虫获取。
a_{13}	F	粉丝男女比例(图表)	用户粉丝的性别，为粉丝属性信息，通过爬虫获取。
a_{14}	F	粉丝关注数量分布(图表)	粉丝关注人数，判断粉丝专一程度，通过爬虫获取。
a_{15}	F	粉丝问题回答数量分布(图表)	粉丝在平台回答的次数，判断粉丝活跃度，通过爬虫获取。
a_{16}	F	粉丝画像(词云)	粉丝个人简介，为用户属性信息，用于归纳粉丝群体特征，通过爬虫获取。

知乎的内容输出主要依靠大 V 回答的问题以及发布的文章，知乎大 V 以其高质量的内容输出著称。任何用户都可以在知乎注册，通过提出问题、回答问题、发表文章等参与学术交流。知乎的所有网络行为，例如注册、发布、点赞、感谢、收藏、浏览记录、评论、关注、私信、订阅专栏等，都被平台记录下来，本文将该记录称之为用户行为数据。用户行为数据越丰富，越能精确刻画用户特征。为了开展实证研究，本文采集到 16 个数据项，涵盖了用户及其粉丝的主要行为数据(见表 1)。

其中 a_1~a_2 是用户的基本信息(Basic Information，下文简写为 B)，a_3~a_4 反映了用户使用平台的积极性(Positivity，下文简写为 P)，a_5 则从互动方面反映了大 V 的影响力情况(Influence，下文简写为 I)，a_7~a_{10} 可以提取描述主题偏好的方向(Preferred direction，下文简写为 D)，a_{11}~a_{15} 通过对比粉丝的画像体现了大 V 的粉丝群体特点(Fan characteristics，下文简写为 F)。为了利用这些数据对高影响力用户进行画像，本文提出了 5 个维度的用户画像模型(Character portrait model，下文简写为 CPM)，即 CPM = {B, P, I, D, F}。

4. 数据分析

(一) 词频统计

我们将上文集成和归约后的数据集进行词频统计，为下文可视化做基础工作。

(二) 情感分析

我们对用户回答的问题进行了情感分析，判断相比较之下是积极还是消极。我们借助 Python 的工具“SnowNLP”来进行情感分析，0 表示消极，1 表示积极，根据 0~1 之间的占比及其分布，可以大致概括用户关注话题的正面与负面情况。

(三) 数据可视化

数据可视化，我们采用了 Python 自带的工具 pyecharts、matplotlib 帮助我们做数据的可视化处理，让我们采集的数据以一个更加直观的方式呈现。

我们从多个维度生成了这些用户的可视化模型。下面以知乎大 V 苏菲为例展示部分维度统计模型及数据解析。

1) 苏菲粉丝的 VIP 用户比例(见图 1)。可知苏菲的粉丝群体全部由非 VIP 用户组成。

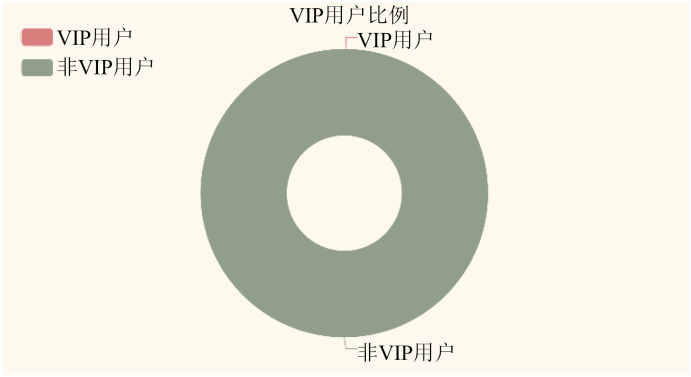


Figure 1. Proportion of VIP users of Sufi fans
图 1. 苏菲粉丝的 VIP 用户比例

2) 苏菲粉丝的男女比例(见图 2)。可以了解到该群体粉丝大部分性别未知，而已知信息的粉丝中男女比例基本持平。

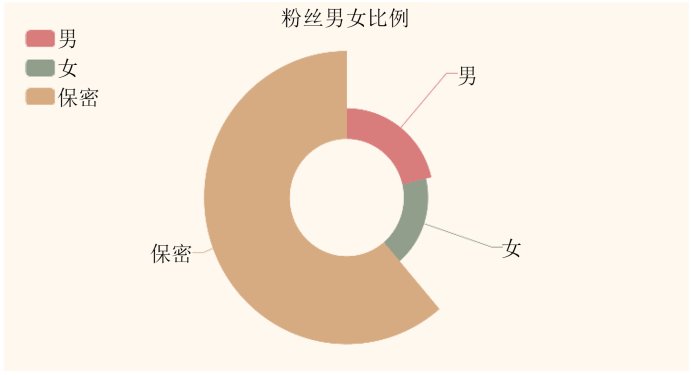


Figure 2. Sex ratio of Sufi fans
图 2. 苏菲粉丝的男女比例

3) 苏菲粉丝的关注数统计(见图 3)。可以得知苏菲 90%的粉丝关注人数少于 10 人，即具有比较专一的粉丝群体。

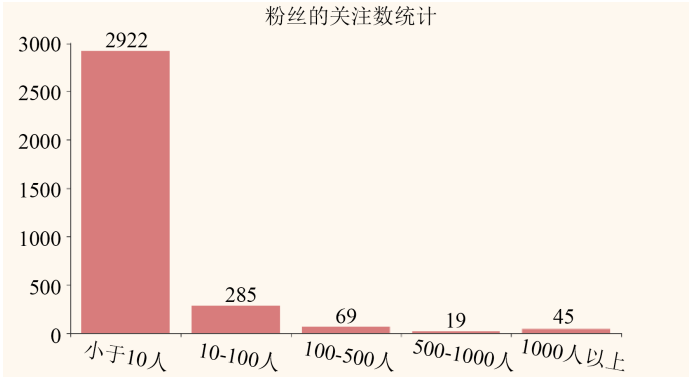


Figure 3. Statistics of the number of Sufi fans' attention
图 3. 苏菲粉丝的关注数统计

4) 苏菲粉丝的问题回答数统计(见图 4)。苏菲大多数粉丝的问题回答数小于 5 次，说明该粉丝群体较少参与答题互动。

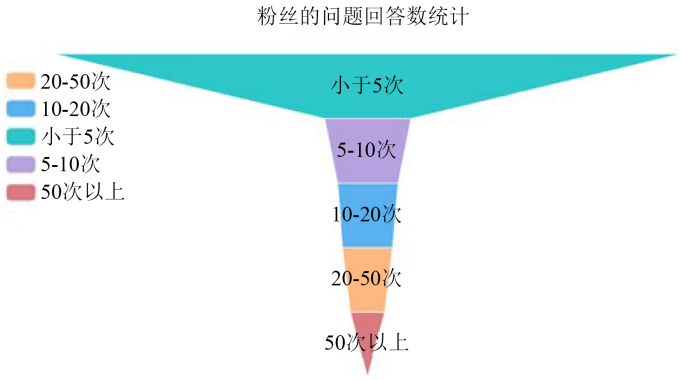


Figure 4. Statistics of Sufi fans' answers to questions
图 4. 苏菲粉丝的问题回答数统计

5) 苏菲粉丝群体的简介关键词词云(见图 5)。可以看到苏菲的粉丝主要由在读大学生和工程师组成。

Figure 5. Introduction key words cloud of Sufi fans
图 5. 苏菲粉丝群体的简介关键词词云

6) 苏菲回答的问题生成的词云(见图 6)。可以看出苏菲主要进行日本文化相关问题的回答互动。

Figure 6. Sophie's word cloud for answering questions
图 6. 苏菲回答问题的词云

7) 苏菲所关注的人的简介关键词词云(见图 7)。由此可以得知苏菲的爱好领域囊括互联网、心理学、公众话题等。

Figure 7. Sufi's profile key words cloud
图 7. 苏菲所关注的人的简介关键词词云

8) 苏菲所关注的人的个性签名的情感分析柱状图(见图 8)。横坐标 0.0 为消极，1.0 为积极，纵轴为数量。可以看出苏菲所关注的人更多具有积极的个性签名或无情感色彩的个性签名。

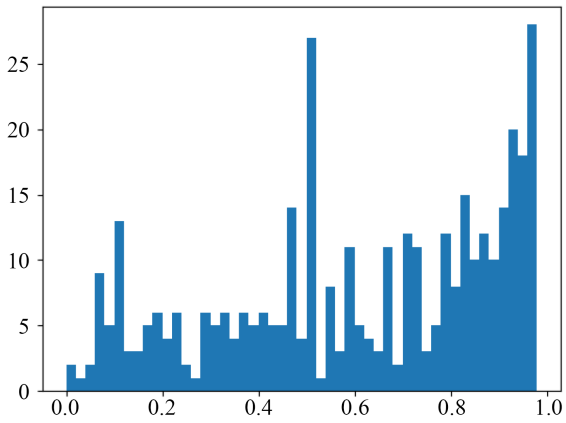


Figure 8. Emotion analysis histogram of Sufi's personal signature
图 8. 苏菲所关注的人的个性签名的情感分析柱状图

9) 苏菲回答的问题情感分析柱状图(见图 9)。横坐标 0.0 为消极，1.0 为积极，纵轴为数量。可以看出苏菲回答的问题更多具有积极的情感色彩，即苏菲更倾向参与正能量互动。

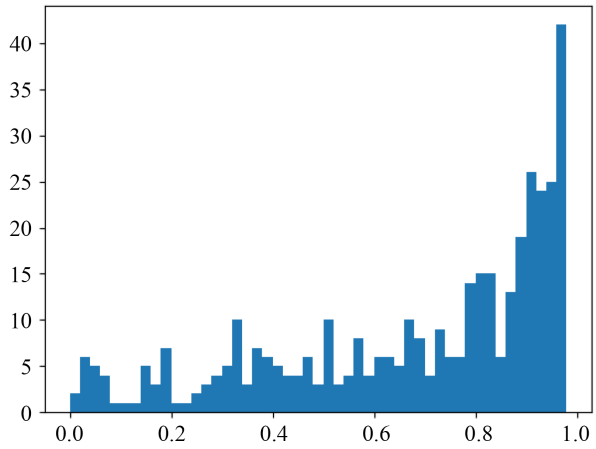


Figure 9. Sophie's answer to the question emotion analysis histogram
图 9. 苏菲回答的问题情感分析柱状图

5. 画像构建

在知乎，高影响力的用户群体主要由两部分组成：

一是已经具有影响力的各行各业的领军人物，这类群体的特点是已经具有一定的粉丝基础。例如：雷军、李开复、周冬雨、郎朗……这些用户大多本身就是各领域的优秀代表或公众人物，其对于一些行业相关问题的回答更具权威性，所以自然能够产生许多优秀的回答，他们在现实生活中也拥有庞大的粉丝数量，因此他们在知乎上也有很高的影响力和关注度，但这类用户的活跃度往往并不是很高。

二是凭借产出高质量回答来逐步提升影响力的答主，在其中有热衷于回答某一类问题各行各业的“专家”们，如叶清波(家具设计话题优秀回答者)、张小北(电影话题优秀回答者)、苏菲(日本文化话题优秀回答者)、KnowYourself (心理学话题优秀回答者)、丁勾(科幻话题优秀回答者)、等等，这些用户通常只喜欢自己擅长领域的话题，凭借独特的回答风格或者高质量的内容往往能够收获不少关注者与赞同，原创性较高，即内容生成积极性高，具备一定的内容创新能力。其中很多还能成为某一行业的优秀回答者，并且这类用户回答问题的活跃度非常高。

除此之外，还有一些用户则是知识面异常广阔，回答问题的相关类型较为丰富，例如：肥肥猫、芝士就是力量等等。同样，这类用户在知乎的活跃度也是非常之高，但没有特定回答的话题。

我们分别选取了这两类用户来做调研，从已有数据库中随机选取了一位公众人物，两位不同话题领域的高质量内容输出用户。接下来我们会从用户画像模型(CPM = {B, P, I, D, F})的 5 个维度来进行具体分析[9]。

(一) 用户基本信息(Basic Information)

Table 2. Comparison of screenshots of high-impact user homepages
表 2. 高影响力用户主页截图对比

苏菲	叶清波	周冬雨
		

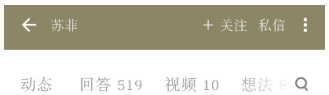
由表 2 用户首页截图可以得知三位用户的基本信息：

- 苏菲的研究领域为日本文化、日本旅游，从事互联网行业；
- 叶清波的研究领域为家居设计，从事创意艺术行业；
- 周冬雨是知名演员，从事演艺行业。

(二) 用户使用平台的积极性(Positivity)

积极性指标(P)与提问数、回答数、点赞收藏数、文章或专栏发表数等知乎用户行为与积极性指标正向相关。

Table 3. Comparison of users' enthusiasm for using the platform
表 3. 用户使用平台积极性对比

	苏菲	叶清波	周冬雨
提问数			
回答数			

从表 3 数据可以得到三位用户的提问数与回答数：

- 苏菲的提问次数 58，回答次数 519，说明苏菲使用平台颇为积极，且回答问题次数更多，从而可以判断影响力较高的原因在于积极回答问题。
- 叶清波的提问次数 15，回答次数 231，说明他使用平台比较积极，但由于回答次数少于苏菲，所以可以推测他的影响力相对于苏菲较弱的原因在于此。
- 周冬雨的提问次数 3，回答次数 24，说明周冬雨使用平台的频度不太积极。这组数据不符合前面推测的结论，所以可以推测周冬雨获得高影响力的原因在于知名人物本身。

(三) 用户的影响力(Influence)

知乎用户的影响力(I)可以通过一定指标来描述，我们基于用户的粉丝数量以及他回答的感谢数、赞同数、收藏数和被评论数等综合分析用户的影响力。

根据我们对目标用户公开信息的爬取与收集，可以获得三位用户的粉丝数量和获赞总数如表 2 所示。

- 苏菲的粉丝数 517,237，关注数 474，获赞次数 762,079。可以看出苏菲具有很多的粉丝和很高的人气，从而判断出这也是她平台影响力较高的原因。
- 叶清波的粉丝数 201,124，关注数 446，获赞次数 89,957。可以看出叶清波虽然具有较多的粉丝和人气，但不如苏菲的粉丝和获赞多，从而可以推测知乎用户中对家居设计领域感兴趣的用户相对于对日本文化感兴趣的用户较少，即该行业相对冷门。
- 周冬雨的粉丝数 417,866，关注数 0，获赞次数 189,921。周冬雨的关注数为 0 而粉丝数目和获赞次数却在高影响力用户中排名领先。从而可以加强上文的推断——周冬雨获得高影响力的原因于知名人物本身。

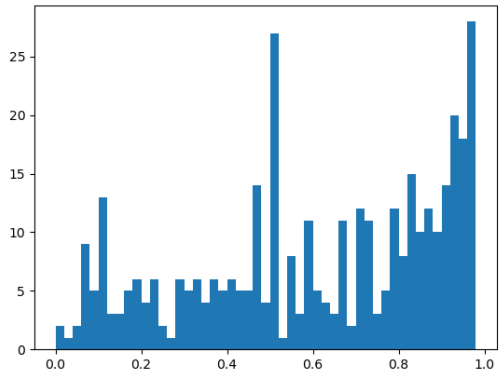
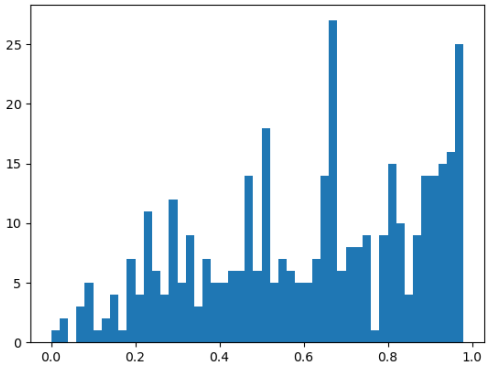
(四) 用户的偏好方向(Preferred Direction)

兴趣是人们活动的巨大动力，用户偏好(P)是他们输出高质量内容时做出的理性的、具有倾向性的选

择，知乎大 V 的兴趣偏好反映了他们关注的领域、话题以及学术方向，本文选择用他们设计的问题中的主题词(关键词)来描述。

1) 用户所关注者的个性签名情感分析

Table 4. Sentiment analysis of personal signatures of users' followers
表 4. 用户所关注者的个性签名情感分析

苏菲	叶清波
	

横坐标 0.0 为消极，1.0 为积极，纵轴为数量，通过表 4 我们可以看出：

- 苏菲所关注的人的个性签名的情感分析柱状图。可以看出苏菲所关注的人更多具有积极的个性签名或无情感色彩的个性签名，从而推测苏菲更倾向于在知乎平台获取积极信息。
- 通过观察叶清波关注的人的个性签名情感分析柱状图可以看出，他们的个签情感大多数是积极色彩的，从而推测出叶清波对信息的需求也大多是正能量的。

2) 用户所关注者的研究领域

Table 5. Word cloud comparison of research areas of users' followers
表 5. 用户所关注者的研究领域词云对比

苏菲	叶清波
	

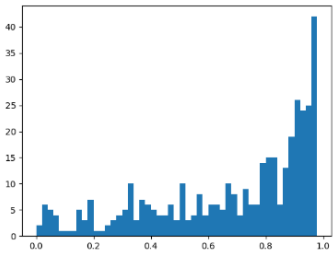
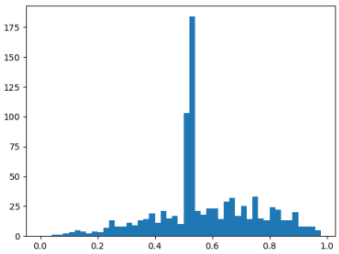
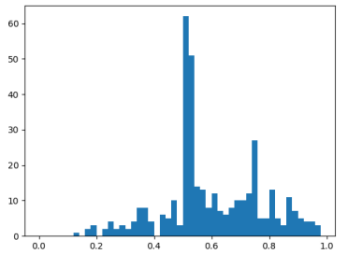
通过表 5 的对比我们可以看出：

- 由苏菲所关注的人的研究领域词云可以得知苏菲的关注领域囊括互联网、心理学、公众话题等，这些领域与她所从事的日本文化领域没有直接关联，由此可以推测苏菲的爱好较为广泛。
- 由叶清波所关注的人的研究领域词云可以得知叶清波的关注领域基本与本领域相吻合，包括室内设

计、建筑设计、家具、装修等等。此外词云中还出现了汽车、美食等领域，由此可以推测叶清波的业余爱好与这些相关。

3) 用户回答问题的情感分析

Table 6. Comparison of sentiment analysis histograms for users to answer questions
表 6. 用户回答问题的情感分析柱状图对比

苏菲	叶清波	周冬雨
		

横坐标 0.0 为消极，1.0 为积极，纵轴为数量，通过图表 6 我们可以直观的看出：

- 由苏菲回答的问题情感分析图，可以看出苏菲回答的问题更多具有积极的情感色彩，即苏菲更倾向参与正能量互动。另外考虑到苏菲从事的行业是跨文化交流、跨国旅游一类，褒义色彩的词汇也对文化的传播、吸引读者具有促进作用。
- 由叶清波回答的问题情感分析图，可以看出叶清波回答的问题大多数不包含感情色彩。原因同上，叶清波从事室内设计行业，回答的问题专业性较强，具有较强的参考借鉴价值，所以感情色彩不太明显。
- 由周冬雨回答的问题情感分析图，可以看出周冬雨的回答包含的感情色彩处于中等偏上。作为公众人物，在公开社区问答平台的言论一般情况下不包含消极词汇，所以她的回答大多数无感情色彩或者具有积极色彩。

4) 用户回答问题的可视化模型分析

Table 7. Word cloud comparison of keywords answered by users
表 7. 用户回答问题的关键词词云对比

苏菲	叶清波	周冬雨
		

通过表 7 我们可以直观的看出：

- 由苏菲回答的问题生成的词云可以看出，苏菲主要进行日本文化、跨国旅游相关问题的回答互动，说明苏菲的回答偏好与本领域对口。
- 由叶清波回答的问题生成的词云可以看出，叶清波主要进行家具设计和获取家居设计信息渠道的指

导相关问题的回答互动，与本人领域对口且有针对性。

- 由周冬雨回答的问题生成的词云可以看出，她主要对自己的影视作品的提问进行了回复，与演员本人相关且阐释了自己在影视剧中的角色。由此推测出这些问题是由粉丝提出的。

(五) 粉丝群体画像(Fan Characteristics)

我们通过对粉丝群体画像(F)可以从另一个角度丰富该大 V 的人物画像，下面我们将从粉丝的 VIP 比例、男女比例、关注数量、问题回答以及一些公开信息来进行分析，从而抽取出该大 V 的特征标签。

1) 粉丝群体 VIP 比例

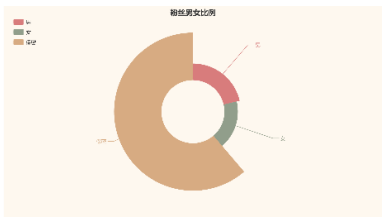
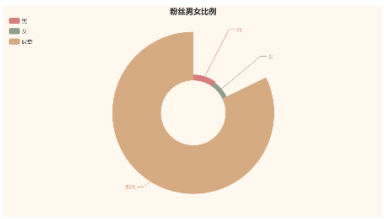
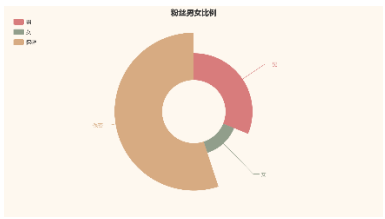
Table 8. Comparison of visual model of fan VIP ratio
表 8. 粉丝 VIP 比例可视化模型对比

苏菲	叶清波	周冬雨
		

由表 8 可见，三位用户的粉丝群体全部由非 VIP 用户构成。由此可以分析到粉丝群体的 VIP 特征与用户无关，而与知乎平台有关。可以推测知乎平台的大部分功能可以免费使用，是一个服务性较强的平台。

2) 粉丝群体男女比例

Table 9. Comparison of visual model of male-female ratio
表 9. 粉丝男女比例可视化模型对比

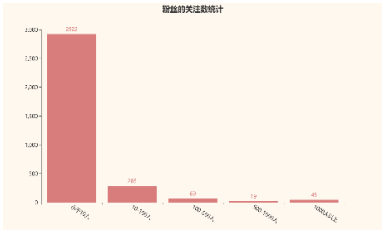
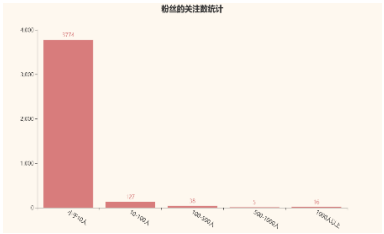
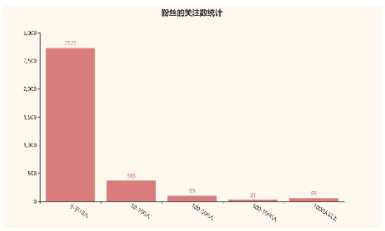
苏菲	叶清波	周冬雨
		

通过表 9 我们可以直观的看出：

- 通过观察苏菲的粉丝群体男女比例可以看出，除去性别保密的粉丝用户，男女比例基本持平。由此可以分析出有了解日本文化或出境旅游需求的群体无性别差异。
- 而观察叶清波的粉丝男女比发现，他的粉丝中性别保密的用户占比很大。除此之外男女比例也是基本持平的。由此可以分析出该粉丝群体更注重相关知识的获取，或者对有针对性别的个性化推荐无过多需求。
- 观察周冬雨的粉丝群体可以发现除去性别保密的粉丝用户，男性粉丝占比更大。说明周冬雨作为女性演员更受男性群体的欢迎。

3) 粉丝关注数量分布区间

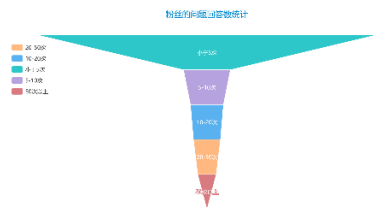
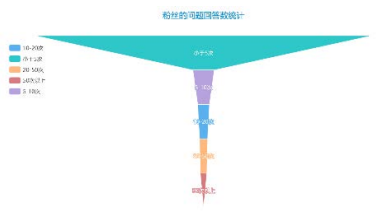
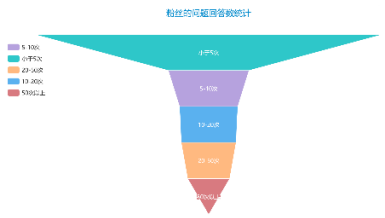
Table 10. Comparison of visualization models for the number of followers
表 10. 粉丝关注数量可视化模型对比

苏菲	叶清波	周冬雨
		

由表 10 可见，三位用户的粉丝关注人数小于 10 人的基本各占 90%。由此可以分析这三位用户粉丝的专一程度都处于较高水平。然而值得注意的点在于叶清波粉丝在关注数量位于 10~100 人这一区间的数量明显少于另外两位用户，该区间以右的各个区间也有所减少。这恰好印证了由表 9 推出的结论，即该粉丝群体对这一领域知识获取的针对性较强，换而言之，该群体用户使用知乎可能主要是为了了解家居设计领域的知识，出于知识的获取目的而不是浏览目的。

4) 粉丝问题回答数量分布

Table 11. Comparison of visualization models for the number of fan questions answered
表 11. 粉丝问题回答数量可视化模型对比

苏菲	叶清波	周冬雨
		

由表 11 可见，三位用户的粉丝回答数量小于 5 的都占有最大比重，说明他们三位的大部分粉丝都不是活跃用户。

- 观察苏菲粉丝的答题数量分布区间可以看到较少粉丝的回答数量多于 10 个，说明她的粉丝群体大部分是非活跃用户。
- 叶清波粉丝的答题数量分布区间显示极少粉丝的回答数量多于 10 个，说明她的粉丝群体绝大多数是非活跃用户。再次印证了上文分析的结论，且延伸为这些粉丝用户是知识的需求方而不是供给方。
- 从周冬雨粉丝的答题数量分布区间来看，即使回答数小于 10 个的占了大多数，但大于 10 个的区间明显多于苏菲和叶清波。说明周冬雨粉丝的活跃程度更高，推测出对娱乐圈感兴趣的用户群体可能整体更加活跃。

5) 粉丝主页关键词可视化模型

由表 12 可见，三位用户的粉丝最主要组成群体都是学生。这一点说明不是个性而是共性特征，可以推断学生是知乎平台的最活跃群体。

Table 12. Comparison of keyword word cloud on fan homepage
表 12. 粉丝主页关键词词云对比

苏菲	叶清波	周冬雨
		

- 观察苏菲的粉丝主页关键词词云可以发现粉丝群体的主要组成除了学生还有设计师和工程师。说明苏菲的回答对这两类人群具有参考价值。
- 通过观察叶清波的粉丝主页关键词词云不难发现设计师基本上占有于学生群体同等的比重，而与设计有关的标签之和则超过了学生的比重。不难说明叶清波的粉丝群体具有较强的集中性。
- 而周冬雨的粉丝群体除去最大比重——学生之外，其他类别的标签就不太明显，比较综合。这一点体现了公众人物的特点，即广泛为各类群体所接受。

6. 结语

本文选取网络问答社区知乎的高影响力用户作为研究对象，从用户基本信息、使用平台的积极性、互动情况、偏好方向及粉丝群体画像 5 个方面构建了 3 位不同领域高质量用户画像。通过对高影响力用户及其粉丝群体画像，试图了解高影响力用户内部的数据特征，并解决了如何有效进行文本特征提取，情感分析的问题。本文将用户画像的概念运用到社会化问答社区的用户分析中，为分析社会化问答社区用户层面甄别提供了帮助。本研究可以作为知乎数据挖掘在商业情报、社会研究、舆情研究等领域的应用。进而扩展为分析用户群体、监控社群行为的基础。

由于数据集的限制，本文没有从文本内容，如话题、文章等更细致的角度分析用户画像。未来笔者会针对文本内容、用户的变化情况做出分析，得到更有意义的社会化问答社区用户画像；并针对社会化问答社区发展的不同阶段进行用户子社区识别、分类以及演变分析。

参考文献

[1] 王凌霄, 沈卓, 李艳. 社会化问答社区用户画像构建[J]. 情报理论与实践, 2018, 41(1): 129-134.

[2] 付少雄, 陈晓宇. 知识网红内容表现力的影响因素分析: 以知乎为例[J]. 情报资料工作, 2019, 40(6): 81-89.

[3] 向世康. 场景式营销: 移动互联网时代的营销方法论[M]. 北京: 北京时代文化书局, 2017.

[4] 刘禹辰. 基于情感分析的 Android 平台用户画像方法研究[D]: [硕士学位论文]. 北京: 北京交通大学, 2019.

[5] 张莉曼, 张向先, 吴雅威, 郭顺利. 基于小数据的社交类学术 App 用户动态画像模型构建研究[J]. 图书情报工作, 2020(5): 50-59.

[6] 魏明珠, 张海涛, 刘雅姝, 徐海玲. 多维属性融合的社交媒体高影响力人物画像研究[J]. 图书情报知识, 2019(5): 73-79+100.

[7] 王飞翔. 数据挖掘技术在问答社区中的应用[D]: [硕士学位论文]. 南京: 南京邮电大学, 2018.

[8] 陈烨, 陈天雨, 董庆兴. 多视角数据驱动的社会化问答平台用户画像构建模型研究[J]. 图书情报知识, 2019(5): 64-72.

[9] 袁润, 王琦. 学术博客用户画像模型构建与实证——以科学网博客为例[J]. 图书情报工作, 2019, 63(22): 13-20.