学术聚焦·社交媒体用户画像

Academic Focus: User Profile of Social Media

【编者按】 用户画像近年来逐渐成为学界关注的热点。作为基于用户真实数据的虚拟代表和目标用户模型，用户画像在精准营销、个性化服务等方面得到日益广泛的应用。与此同时，关于图书馆用户画像的研究也开始兴起，主要涉及图书馆智能化、精准化的信息推送和推荐服务等，这与图书馆一直以来重视用户研究的传统有着密切的关系。

大数据环境下，社交媒体上蕴含的海量数据为用户信息标签化提供了蓝本。可以预见，社交媒体将成为用户画像研究的又一个新场景。

在本期里，我们新开设了“学术聚焦”栏目，并把第一期的主题聚焦在“社交媒体用户画像”上，分别是《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》《多维属性融合的社交媒体高影响力人物画像研究》。我们希望通过这个新设栏目，吸引和聚集更多学术研究中的新观点和新视角。热切期待您的参与。

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

Research on the Model Construction of Multi-View-Data-Driven User Profile for Social Q&A Platform

陈 烨 陈天雨 董庆兴

(华中师范大学信息管理学院，武汉，430079)

[**摘要**] [**目的/意义**]聚焦社会化问答平台，探索多视角数据驱动的用户画像构建框架和方法，旨在更全面、准确地理解用户，进而为用户提供更优质、精准的信息服务。[**研究设计/方法**]根据社会化问答平台用户数据的特点，从数据挖掘和本体论的视角厘清用户数据与用户之间的对应关系，在此基础上提出多视角数据驱动的社会化问答平台用户画像构建模型。[**结论/发现**]该模型包括了用户数据获取、属性沙盒搭建、用户画像实现和用户画像应用等环节，在用户画像实现过程中主要涉及用户画像生成与用户画像更新两个关键环节。[**创新/价值**]建立了多视角用户数据与用户之间的对应关系，阐述了社会化问答平台用户画像的构建框架与方法。

[**关键词**] 用户画像 构建模型 多视角数据 社会化问答平台 生成模型 更新模型

[**中图分类号**] G203 [**文献标识码**] A [**文章编号**] 1003-2797(2019)05-0064-09 **DOI**:10.13366/j.dik.2019.05.064

[**Abstract**] [Purpose/Significance] This study aims to explore the framework and methods of multi-view-data-driven user profile construction for social Q&A platform. It intends to provide a more comprehensive and more accurate understanding of users for service providers, which will help them improve the quality and accuracy of their services. [Design/Methodology] According to the characteristics of user data from social Q&A platform, the corresponding relationship between user data

[**基金项目**] 本文系国家自然科学基金青年项目“基于多视角学习的社会化问答平台用户画像研究”(71904057)、中央高校基本科研业务费专项资金科研项目(CCNU18XJ025)的研究成果之一。

[**通讯作者**] 陈烨(ORCID:0000-0002-7619-3246)，博士，讲师，研究方向：数据挖掘与用户研究，Email：cheny@mail.ccnu.edu.cn。

[**作者简介**] 陈天雨(ORCID:0000-0001-5728-9616)，本科生，研究方向：数据挖掘，Email：tianyuchen@mails.ccnu.edu.cn；董庆兴(ORCID:0000-0003-3512-9333)，博士，副教授，研究方向：知识管理与智能决策，Email：qxdong@mail.ccnu.edu.cn。

composition and user attribute extraction has been clarified from the perspectives of data mining and ontology. Then a multi-view-data-driven model of user profile construction for social Q&A platform has been proposed. [Findings/Conclusion] The model consists of four parts, which are user data acquisition, attribute sandbox construction, user profile implementation and user profile application. And the part of user profile implementation mainly involves two key steps, user profile generation and user profile update. [Originality/Value] The correspondence between multi-view user data and users has been established, and the framework and methods of user profile construction for social Q&A platform have been expounded.

[**Keywords**] User profile; Model construction; Multi-view data; Social Q&A platform; Model generation; Model update

社会化问答平台为用户提供了通过自然语言表达信息需求和通过用户互动满足信息需求的社区，成为网民获取信息的重要途径之一[1]。截止 2019 年，美国著名社会问答网站 Quora 的月独立用户访问量超过 3 亿；而国内著名社会问答网站知乎的日活跃用户达到 2600 万，累计用户已经突破 3 亿，其中付费用户超过 600 万[2,3]。用户在使用社会化问答平台的过程中产生了大量来源多样、内容丰富、形态各异的用户数据。这些用户数据呈现出多态性、多源性和异构性的特点，是一类典型的多视角数据，为社会化问答平台服务方全面、准确地理解用户提供了丰富的数据来源。但如何组织、管理和利用多视角用户数据使之服务于产品设计与服务优化是社会化问答平台用户研究中重点关注的问题。用户画像利用各类用户数据实现用户属性特征的揭示和组织，提供了一种全面立体刻画用户的框架和工具，为社会化问答平台用户研究提供了新思路。基于此，本文从社会化问答平台多视角用户数据出发，在建立多视角用户数据与用户的对应关系的基础上，提出社会化问答平台用户画像构建模型，并对社会化问答平台用户画像的应用场景进行分析，旨在为社会化问答平台用户研究提供一套较为完整的研究与实践的框架与思路。

1 相关研究

1.1 用户画像研究概述

“用户画像（user profile）”是一个新兴的概念，兴起于互联网行业。用户画像提出之初旨在通过给用户“打标签”的方式标识用户特征，并基于用户标签实现用户分类管理。随着用户画像在精准营销、个性化服务等应用实践中发挥出越来越大的价值，学界关于用户画像的内涵和外延、实现流程与模型方法的探讨也逐渐深入，历经初始阶段和起步阶段，于 2015 年进入发展阶段，主要涉及计算机、经济金融、图书情报、新闻传媒、工业技术等领域[4]。计算机领域的研究者认为用户画像是推断用户特征的过程、手段和方法，是为每个用户贴上精确标签的有效手段[5]；统计学领域的研究者将用户画像理解为对现实世界用户的数学建模[6]；心理学领域的研究者认为用户画像是对用户特征的勾画，将用户的特点直观明了地表现，反映用户的触点和痛点，达到与产品和服务的链接甚至是评价[7]；图书情报领域的研究者则将用户画像理解为用户信息标签化，是建立在一系列数据之上的目标用户模型[8]。

除了对用户画像的概念进行分析与界定之外，研究者们主要围绕用户画像的实现展开了研究。用户画像的类型多样，从指代对象的角度可以分为个体用户画像和群体用户画像[9]，从应用场景的角度可以分为移动用户画像[10]、医疗用户画像[11]、电力用户画像[12]等。不同类型用户画像的实现流程与方法各有侧重，但综合看来，可以将用户画像的实现流程归纳为数据收集与处理、用户属性划分、用户特征挖掘和用户画像表示等四个主要环节。

用户画像的用户数据来源和类型各异，用户数据的获取方式可以分为直接获取和间接获取。由于直接获取方式效率远远低于间接获取方式，单纯利用直接获取方式已经无法满足新网络环境下用户数据获取的要求，此外，研究表明采用间接获取方式或混合获取方式生成的用户模型精度大于单独采用直接获取方式[13,14]，因此，用户画像以间接获取方式为主进行用户数据采集。用户画像将用户看作具有不同维度属性的对象，通过解剖用户在不同属性上的特征，揭示用户在不同方面的表现和特点，进而有针对性地提供产品或服务。由于应用场景的差异，用户

属性维度划分不尽相同。例如，郭光明[15]将用户画像的目标属性分为事实性属性和行为性属性；刘蓓琳和张琪[16]通过总结具有代表性的用户画像研究，得到目前受到较多关注的六类用户属性维度，分别为基本属性、社交属性、行为特征、兴趣属性、能力属性、心理属性。由于数据来源和应用需要各不相同，在进行用户属性特征挖掘时采用的方法也存在较大差异。目前用户画像研究中的用户属性特征挖掘方法主要包括基于统计的方法[17]、基于数据挖掘的方法[18]、基于机器学习的方法[19]和其他方法[20]。用户画像表示指的是用户特征可视化。用户画像通常以标签的形式表示用户特征，但它有别于直接利用用户生成标签表示用户特征的方法[21,22]，主要从用户数据中挖掘用户不同侧面的特征，并抽象为用户易理解或计算机可读的标签[23]。用户画像标签的内容多样，可以是词汇、短语或概念等，标签的可视化方式可以是向量、描述图表和标签云等。

1.2 社会化问答平台用户画像研究现状

社会化问答平台是在 web2.0 时代背景下发展起来的知识共享平台，以“问答”为主要形式，通过用户、话题与问题之间的连接满足用户的信息需求，具有交互性、共享性、社会化等特点[24]。国内具有代表性的社会化问答平台有悟空问答、知乎、百度知道等，而国外具有代表性的社会化问答平台有 Yahoo! Answers、Quora 等。社会化问答平台的发展伴随着用户与用户之间、用户与信息之间关系的不断发展，但如何利用社会化问答平台中产生的多视角、碎片化用户数据来提升用户体验成为社会化问答平台发展的一大挑战。以全面立体刻画用户为主旨的用户画像为网络用户研究提供了新的研究思路，研究者们尝试将用户画像引入社会化问答平台用户研究中，并着重在构建框架和构建方法方面进行了探索。

在构建框架方面，张海涛等[25]提出从用户需求、用户角色和用户行为等维度构建用户画像概念模型，并遵循数据源确定、典型用户选取、模型细分维度与用户细分标签映射的步骤，实现在线健康社区用户画像；也有研究者利用在线评论数据，首先从用户信息属性、酒店信息属性和用户评价信息属性三个维度构建用户画像概念模型，然后基于本体实现用户画像模型，最后对用户画像进行多维可视化分析[26]；此外，王凌霄等[27]分别从用户资历、用户参与度、用户回答质量以及用户发展趋势四个方面，遵循数据获取、指标设定、特征获取、画像表示的流程，实现社会化问答社区用户画像。尽管面向不同的场景和目标，用户画像构建框架存在区别，但可以将用户画像构建流程作进一步的归纳为用户数据获取、用户特征识别、用户画像表示等三个环节。

在构建方法方面，研究者们主要探索采用何种方法从用户数据中挖掘用户特征。研究过程中，有利用统计分析的方法综合各项用户数据，实现用户特征揭示[28,29]；有基于概念格的方法，用概念标签表示用户特征，建立用户与概念格的映射关系，并通过关联规则挖掘用户特征之间的关联[30]；也有在定义类、对象属性、数据属性、约束条件构建酒店用户画像本体的基础上，实现基于本体的用户画像实例化[31]。

1.3 研究述评

目前关于社会化问答平台用户画像的研究已有一些实践探索，但大多聚焦某一典型社会化问答平台或社会化问答平台的某一具体话题的用户，尚未从较为宏观的角度总结归纳社会化问答平台用户画像的构建及应用问题。此外，目前的研究大多分别从不同类别的用户数据中挖掘用户属性特征，尚未充分利用不同类别用户数据、捕捉不同类别用户数据之间的关联，进而挖掘用户属性特征。基于此，本文根据多视角用户数据的特点，将多视角学习方法引入到多视角用户数据的处理、加工与利用中，并在多视角用户数据分析、用户属性分析以及多视角用户数据与用户对应关系分析的基础上，从较为宏观的角度提出涵盖用户画像生成、更新与应用的全生命期社会化问答平台用户画像构建模型，为社会化问答平台用户画像研究与应用提供参考。

2 多视角用户数据与用户

在构建社会化问答平台用户画像模型之前，需要对社会化问答平台多视角用户数据的特点、用户属性的类别以及多视角用户数据与用户的对应关系进行梳理与分析。

2.1 多视角用户数据

多视角数据（multi-view data）是针对同一对象从不同途径或不同层面获得的特征数据，其呈现出

多态性、多源性、多描述性和高维异构性等特点[32]。社会化问答平台中的用户数据指的是用户在使用社会化问答平台过程中产生的所有数据，包括问题、答案、评论等用户生成内容，查询日志、行为记录、位置标签等用户行为数据等。这些用户数据由不同途径或层面的特征数据构成，来源多样、内容丰富、形态各异，是一类典型的多视角数据。

尽管不同类型的社会化问答平台由于功能、定位的差异，用户数据的内容存在差别，但由于它们的核心功能是“问答”和“社交”。因而，围绕社会化问答平台的核心功能，可以将社会化问答平台多视角用户数据归纳为三个类别：用户基本信息、用户行为数据和用户贡献内容（如表 1 所示）。

用户基本信息指由用户主动提供的关于用户自身情况的描述信息，包括用户昵称、自我描述（如知乎中的“一句话介绍”）、居住信息、教育信息、工作信息等。用户基本信息主要来自客户端的用户主页，内容的呈现形态可能是文字、图片、超链接等。用户行为数据则包括两类，一是可公开获取的用户行为数据，如用户提出问题、回答问题、关注主题、关注用户以及发布文章等行为数据；二是仅后台存储的用户行为数据，如用户日志、操作记录等。可公开获取的用户行为数据来源主要是客户端主页，如特定主题下的问题页面、关注者页面以及用户主页等，内容的呈现形式包括日志、表格等。用户贡献内容指的是用户生成的对平台内容有贡献的内容，例如用户提出的问题、回答的内容、撰写的文章等，也主要来自于客户端主页，内容的呈现形式包括文本、图片、视频、音频、超链接等。

表 1　社会化问答平台多视角用户数据类别

类别	内容	来源	类型
用户基本信息	用户主动提供的自我描述信息	客户端主页（为主）	文本、图像、超链接等
用户行为数据	可公开获取和仅后台存储的用户行为数据	客户端主页后台服务器	日志文件等
用户贡献内容	用户生成的对平台内容有贡献的内容	客户端主页（为主）	文本、图片、视频、音频、超链接等

2.2　用户属性类别

社会化问答平台用户具有某些共同的属性，且每类或每位用户在同一属性上的表现可能各不相同。用户属性是用户属性特征的抽象和概括，用户属性特征是用户属性的具象化表达。从用户属性的角度观察用户，能够抓住用户的基本性质和共同性质，为深入理解用户提供一个统一的、全局的视角。

马克思主义人性观认为，人有两种属性，一是人的自然属性，二是人的社会属性。人的自然属性是指人的肉体存在及其特征；而人的社会属性是指在实践活动基础上人与人之间发生的各种关系及其特征。社会化问答平台中的用户是物理世界中真实存在的个体的虚拟化身，同样拥有人的两种基本属性，即自然属性和社会属性。因此，本文将社会化问答平台用户的属性划分为两大类：用户自然属性和用户社会属性。然而，信息空间中的用户的自然属性和社会属性并不与物理世界中的人的自然属性和社会属性一一对应，也不是人的自然属性和社会属性的简单映射，社会化问答平台中用户的属性具有新的含义和表现形式。

用户的自然属性包括了与用户基本情况相关的用户属性，例如，用户名称、用户性别、用户年龄、用户描述、职业经历、教育背景以及所在地域等。用户的社会属性则指用户在社会化问答平台中与其他用户发生的各种关系及其特征，例如，兴趣属性、社交属性、能力属性等。其中，兴趣属性包括了与用户兴趣（信息需求）相关的用户属性。由于用户的兴趣或信息需求存在不同的状态，包括意识到并已表达、意识到但未表达以及尚未意识到的兴趣，因此可以将用户兴趣属性类别分为用户显性兴趣和隐性兴趣。用户的显性兴趣指被用户已表达的兴趣或需求；用户的隐性兴趣指潜在的、被唤醒的或被认识但未表达的用户兴趣或需求。社交属性主要包括与用户参与平台社交活动相关的用户属性，如参与方式、参与程度等。用户参与方式指的是用户参与平台活动时所扮演的角色，可以从不同角度进行划分。例如，用户提出问题、关注问题或主题时意味着用户希望获得某些具体信息或某些方面的知识，是信息（知识）的需求者；用户回答问题时是输出信息或知识的

过程，是信息（知识）的提供者；用户对其他用户输出的内容进行评价时（包括点赞或差评）是表达认同、赞赏或反对、批评的过程，扮演着信息（知识）的审查者。 此外，用户参与关注互动时扮演的角色可能是关注者或被关注者；参与评论互动时扮演的角色可能是评论者或回复者。 用户参与程度体现在用户关注的人数、主题数、问题数，以及提出、回答的问题数等方面。 能力属性包括了与用户输出能力相关的用户属性。 输出能力指的是用户输出高质量内容的能力，包括提出优质问题的能力、提供优质答案的能力、提供建设性评论的能力等。 用户的能力属性体现在用户的基本信息和行为信息上。 例如，受教育程度高的用户或是某一领域的从业者在其擅长的领域提供优质答案的可能性较高，获得赞数较高的答案可能较好地回答了用户的问题或是获得了其他用户的认同。

2.3 多视角用户数据与用户的对应关系

社会化问答平台提供了一个有边界的信息空间，用户使用社会化问答平台过程中的所有痕迹，即多视角用户数据，是真实用户个性、特点的反映。 通过多视角用户数据，可以洞察真实用户的喜好、特征。

从数据挖掘角度观察多视角用户数据，多视角用户数据具有不同的特征（feature），例如，用户名、职业、提出的问题、回答的问题、评论数等。 这些特征又可以归纳为不同的视角（view），一个视角可以包含单个或多个特征。 例如，将用户名特征和职业特征归纳为用户基本信息视角，将提出的问题和回答的问题归纳为用户贡献内容视角。 而每位用户是一个实例（instance），通常由特征向量（feature vector）表示，而特征向量由特征值（feature value）或特征集合（feature set）构成，例如，U＝{‘某 a’，‘学生’，Q，A，‘30’}，其中‘某 a’表示用户名，‘学生’表示职业，Q 表示用户提出的问题集合（由问题列表构成），A 表示用户提供的答案集合（由答案列表构成），‘30’表示评论数。

从本体论的角度观察用户，用户具有不同的属性（attribute），例如，自然属性、兴趣属性、能力属性等。 每个属性有不同的特征，称为属性特征（attribute feature），例如，如果某用户对电影、综艺话题较感兴趣，那么该用户在兴趣属性上的属性特征为电影、综艺。

将数据挖掘视角下的多视角用户数据与本体论视角下的用户的相关概念和对应关系进行总结（如图 1 所示）。 首先，多视角用户数据具有多个特征，特征（集）构成不同的视角，每项特征对应相应的特征值；其次，用户具有多个属性，每个属性对应相应的属性特征。 其中，多视角用户数据与用户、特征（集）与属性、特征值与属性特征之间存在映射关系，但并不是简单的一一对应的关系。 探究多视角用户数据与用户的对应关系，事实上是为了明确特征（集）与属性的对应关系，即不同视角用户数据与用户属性之间的对应关系。 具体地说，可以将它们的对应关系归纳为一对一、一对多和多对多三种模式。一对一模式指的是一项数据特征对应一个用户属性。一对多模式指的是一项数据特征对应多个用户属性或多项数据特征对应一个用户属性。 多对多模式指的是多项数据特征对应多个用户属性。

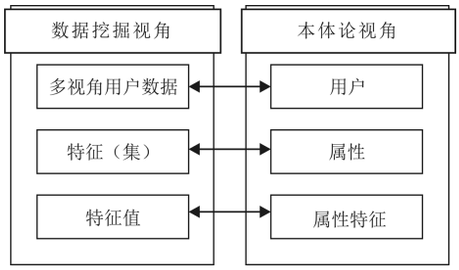

图 1 多视角用户数据与用户的对应关系

在总结归纳多视角用户数据特点、用户属性类别的基础上，建立多视角用户数据与用户的对应关系，明确了多视角用户数据在揭示用户属性特征过程中的重要性。 然而，要通过多视角用户数据揭示社会化问答平台用户属性特征，需要进一步构建服务于应用开发与决策支持的多视角用户数据利用框架，实现从多视角用户数据到用户属性特征再到服务提供与决策支持的贯通。

3 多视角数据驱动的社会化问答平台用户画像构建模型

社会化问答平台用户画像是社会化问答平台用户属性特征的集合，具有多样性和动态性。 用户画像多样性体现在两个方面：目标用户多样性和用户分面多样性。 目标用户多样性指的是既可以面向某些用户群

体生成用户画像，也可以面向单个用户生成用户画像。 用户分面多样性指的是不同用户画像所描绘的用户属性特征有所侧重，可能是用户一个分面或多个不同分面的特征。 用户画像动态性是由用户数据动态性决定的，用户数据在数量、形式和内容维度上都随时间动态变化，因此用户属性特征会随时间发生变化，相应地，社会化问答平台用户画像也具有动态性。

社会化问答平台用户画像的应用场景千变万化，如若将用户所有的属性特征综合到一个用户画像中，一方面在数据融合时需要花费极高的成本；另一方面，在用户画像更新与维护时需要付出较大的代价，且灵活性差。 可行性、实用性更强的解决方案是根据应用需要生成反映用户某一或某些分面特征的用户画像。 基于此，本文提出多数据驱动的社会化问答平台用户画像构建模型。

该模型首先从不同层面、渠道获取多视角用户数据；其次，通过搭建属性沙盒建立多视角用户数据与用户的对应关系，实现多视角用户数据的分流与管理；然后，根据用户画像的应用需要生成并更新用户画像，其中用户画像生成与用户画像更新是构建用户画像过程中最为关键的环节（如图 2 所示）。

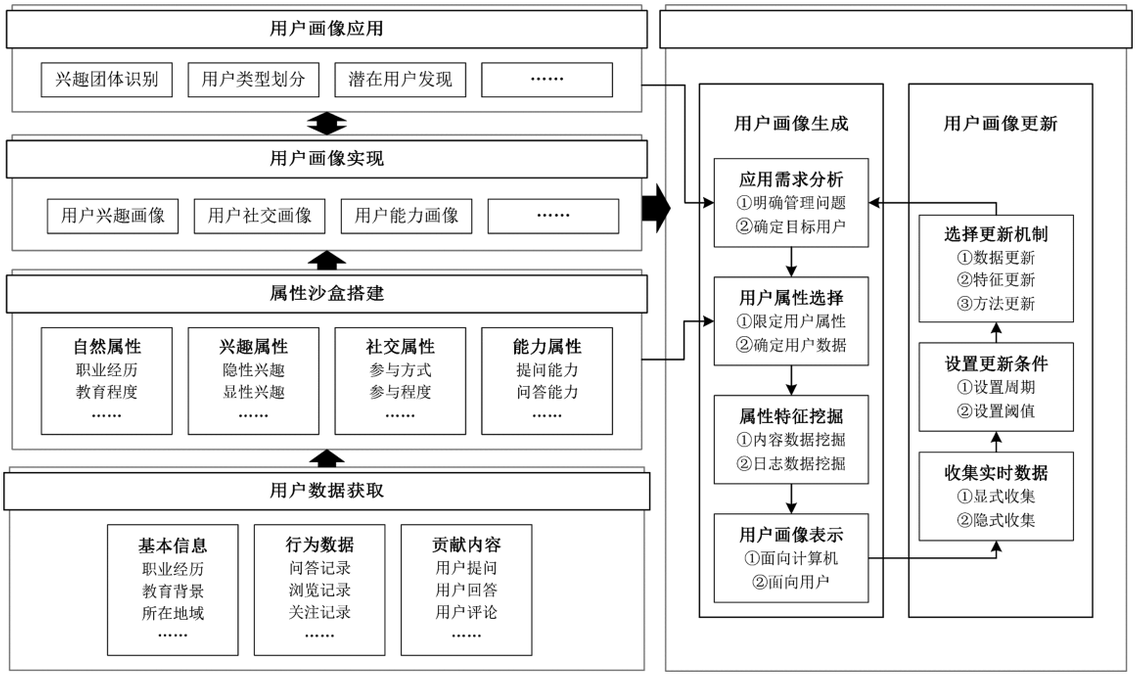

图 2　多视角数据驱动的社会化问答平台用户画像构建模型

3.1　用户画像生成模型

用户画像生成是揭示用户属性特征的过程。 在获取多视角用户数据和构建属性沙盒的前提下，结合用户画像应用需要，可以进一步将用户画像生成过程归纳为应用需求分析、用户属性选择、属性特征挖掘、用户画像表示等四个环节。

（1）应用需求分析。 用户画像服务于应用，为管理问题提供决策支持和判断依据，因此，需求分析是用户画像生成的第一个环节。 需求分析环节应该解决的问题包括两个方面：第一，确定将利用用户画像解决什么管理决策问题？ 社会化问答平台中面临的管理决策问题包括了内容管理、用户管理和技术管理三个方面，其中内容管理涉及内容过滤、质量监督、舆情监控等；用户管理涉及用户激励、用户引导、用户分类、用户反馈等；技术管理指的是通过技术手段辅助内容管理与用户管理。 第二，确定涉及这一问题的相关目标用户。 不同的管理决策问题所涉及的目标用户群体存在差异，而面向单个用户或群体用户构建用户画像的侧重点有所差异。 因此，在

需求分析环节需要明确这一个问题。

（2）用户属性选择。 属性选择环节则针对上一环节提出的管理决策问题和目标用户，选择与之相关的用户属性及其对应的多视角用户数据。 因此，该环节解决的问题是利用哪些用户属性及其对应的多视角用户数据生成用户画像？ 在选择用户属性时，通常采用“按需获取”的策略，根据应用的需要选择一种或多种用户属性及其对应的多视角用户数据用于生成用户画像。

（3）属性特征挖掘。 选取用户属性及其对应的多视角用户数据后，需要根据多视角用户数据的类型、内容、数量特点以及目标进行用户属性特征挖掘。 针对不同的用户数据将选取不同的用户属性特征挖掘方法。 表 1 中总结了社会化问答平台的用户数据类型包括文本、图像、视频、音频、超链接、日志等，可以将其进一步归纳为内容数据（文本、图像、音频、视频）、结构数据（超链接）和日志数据（日志）。 内容数据和日志数据是揭示用户属性特征的最重要数据来源；结构数据体现了网页之间的链接关系，由服务提供方定义，对用户属性特征的揭示较为有限。 针对内容数据，通常采用自然语言处理、统计分析、文本挖掘和机器学习、社会网络分析的方法，具体地有回归分析、相关分析、分类聚类、主题挖掘、特征抽取、特征融合等。 针对日志数据，有专门的日志挖掘方法可供用户属性特征挖掘。 此外，由于多视角用户数据之间存在某些特定的相关关系，分别针对某个特征数据进行用户属性特征挖掘具有局限性，因而，将多视角学习方法引入用户属性特征挖掘，既最大化多个视角数据之间的一致性，又强调每个视角数据的独特性，能够更加全面而准确地理解用户。

（4）用户画像表示。 用户画像表示指的是用户属性特征表示，通常以标签的形式表示用户属性特征。 标签具有概括性，凝练了用户属性特征中的关键信息，但标签的内容和形式多样，既可以是易于计算机处理的特征向量，也可以是便于使用人员理解的短语、图片或图标等。

3.2 用户画像更新模型

由于用户具有背景、需求、角色、行为多样性和动态性，在利用用户画像的过程中，也需将时间因素考虑在内，对用户画像进行更新。 用户画像更新主要涉及三个方面的问题：一是，如何获取实时变化的用户数据？ 二是，如何设置合适的用户画像更新触发条件？ 三是，选择何种高效的用户画像更新机制？

（1）收集实时用户数据。 用户数据的收集方式可以分为显式收集和隐式收集。 显式收集指的是用户直接参与数据收集工作，例如邀请用户填写问卷调查表、情况调查表等；隐式收集指的是用户间接参与数据收集工作，例如通过网络爬虫爬取平台上的用户数据，或是通过 API 获取用户日志文件等。 对于相对稳定的用户数据，如用户基本情况数据，可采取显式收集方式，对于变化较快的其他用户数据，通常采用隐式收集方法。

（2）设置更新触发条件。 用户画像更新的触发条件有两种设置方式，分别是设置更新周期和设置更新阈值。 无论是设置更新周期还是更新阈值，都需要在掌握用户属性特征随时间变化的规律与特点的前提下展开。 分析用户属性特征动态性时，可以从用户数据的内容、形式和数量等维度，利用时序分析和比较分析等方法掌握用户属性特征随时间变化的规律；此外，由于用户画像的生成过程是用户属性特征标签的生成过程，涉及二分类和阈值分类问题，可以借助相应的模型性能评价指标探索模型随时间变化的特征与规律。 如果用户属性特征随时间的变化呈现出明显的周期性，通常采用设定更新周期的方式；否则，采用设定更新阈值的方式。

（3）选择画像更新机制。 根据用户画像对数据的内容、形式和数量的敏感度，确定是否需要更新用户数据、特征体系或训练方法。 如若需要更新用户数据，则可以采取完全更新或增量更新的策略，其中，完全更新指读取当前时间戳所有历史用户数据重新生成用户画像，增量更新指获取当前时间戳与上一个时间戳之间用户数据，只更新发生变化的部分。 如果需要更新特征体系，则涉及多视角数据融合，可以采用协同训练、多核学习和子空间学习等多视角学习方法实现多视角数据融合。 而无论是用户数据更新还是特征体系更新，都可能涉及训练方法的更新。

4 社会化问答平台用户画像应用场景

随着社会化问答平台在人们获取信息、知识和社

会支持的过程中扮演着越来越重要的角色，如何进一步提升服务的质量、改善用户体验成为服务提供方的工作重心。 而在提供服务的过程中，一些问题仍未得到很好地解决，包括如何准确地捕捉用户兴趣，提供精准的服务；如何面向不同社交偏好的用户，营造多元化的社交氛围；如何基于多视角用户数据，提供个性化的内容等。 多视角数据驱动的社会化问答平台用户画像能够在一定程度上解决上述问题。

用户兴趣会随着用户所处的工作、生活状态的变化而发生变化，可以分为长期兴趣和短期兴趣，如何识别用户长、短期兴趣是为用户提供精准服务的前提之一。 通过构建用户画像来描述用户兴趣，可以捕捉用户兴趣随时间动态变化的特点，实现用户兴趣的动态跟踪，帮助识别用户的长期兴趣和短期兴趣，并在此基础上有针对性地提供信息服务，从而提高信息服务的精准度。

用户由于受到需求、习惯、价值观、技能等因素影响，更倾向于浏览自己感兴趣的内容，具有选择性心理和注意性理解的行为特点，久而久之就有可能陷入“信息茧房”之中。 客观上“信息茧房”的产生体现了以用户需求为中心的服务理念以及去中心化的内容生产模式，但其带来的消极影响也不容忽视，具体表现为社会黏性降低、群体意见极化、关注内容单一等。 针对这些问题，可以借助社会化问答平台用户画像寻找解决方案。 一方面，通过计算用户画像相似度，确定与目标用户在兴趣上具有强关联的优质用户，为相似度较高的用户群体之间搭建沟通渠道，拓展用户的信息源，增加“信息偶遇”的机会，打破由于“信息茧房”带来的用户区隔。 另一方面，基于多视角用户数据构建用户画像，建立不同视角用户数据之间的关联，更全面、更立体地捕捉用户属性特征，发现用户的潜在兴趣或需求，提升社会化问答平台推送内容的精准度。

为了优化问答呈现结构，满足用户个性化的信息需求，可以通过用户画像刻画用户评价、回答等能力属性，对问答的回答者进行整合与筛选，依据不同用户关注的问题领域及能力属性细粒度匹配相应问题；同时可以将用户对社会化问答平台的贡献进行分级，如获赞数、发表文章数、回答数量、被关注及收藏数量等多方面，甄别优劣用户，并以此为依据过滤低质量、庸俗化的答案，优化平台价值，提高用户黏性。

5 总结与展望

多视角用户数据为社会化问答平台用户研究提供了有力的数据支撑，而用户画像为社会化问答平台用户研究提供了一个新的思路和视角。 基于此，本文在分析社会化问答平台多视角用户数据特点、用户属性类别以及多视角数据与用户对应关系的基础上，提出了包括用户数据获取、属性沙盒搭建、用户画像实现和用户画像应用等环节的多视角数据驱动的社会化问答平台用户画像构建模型，该模型为利用多视角用户数据、实现全面、精准地刻画用户提供了一个较为完整的框架。 从当前社会化问答平台服务中的应用需求出发，社会化问答平台用户画像可以在用户兴趣动态捕捉、用户特征深度挖掘、平台内容个性化提供等方面发挥重要的作用。

目前，关于社会化问答平台用户画像的研究仍处于实践与理论探索阶段，需要在已有研究基础上，进一步探索冷启动情境下的用户画像生成问题，面向具体应用场景的用户画像实现与利用问题，以及面向多元应用场景的用户画像融合与利用问题等。

作者贡献说明

陈　烨：提出研究思路，设计研究方案，撰写论文与修订；
陈天雨：收集和梳理文献，撰写部分论文；
董庆兴：撰写部分论文，论文最终版本修订。

参考文献

1 Shah C, Oh S, Oh J S. Research Agenda for Social Q&A[J]. Library & Information Science Research, 2009, 31(4): 205-209.

2 DMR. 12 Interesting Quora Statistics and Facts(2019)[EB/OL]. [2019-02-20]. https://expandedramblings.com/index.php/quora-statistics/, 2019a.

3 DMR. Interesting Zhihu Statistics and Facts(2019)[EB/OL]. [2019-02-20]. https://expandedramblings.com/index.php/zhihu-statistics-and-facts/, 2019b.

4 吴加琪. 我国用户画像研究的知识网络与热点领域分析[J]. 现代情报, 2018, 38 (8):132-137,145.

5,19 马超. 基于主题模型的社交网络用户画像分析方法[D]. 合肥：中国科学技术大学, 2017.

6 李映坤. 大数据背景下用户画像的统计方法实践研究[D]. 北京：首都经济贸易大学,2016.

7 饶璇. 基于留存与流失用户画像提升用户研究的效果[D]. 武汉：华中师范大学, 2017.

8,17,27,28 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.
9 张哲. 基于微博数据的用户画像系统的设计与实现[D]. 武汉：华中科技大学,2015.
10 黄文彬，徐山川，吴家辉等. 移动用户画像构建研究[J]. 现代情报，2016，36(10)：54-61.
11 王智囊. 基于用户画像的医疗信息精准推荐的研究[D]. 成都：电子科技大学,2016.
12 孟巍，吴雪霞，李静，等. 基于大数据技术的电力用户画像[J]. 电信科学，2017，(S1)：15-20.
13 Quiroga L M，Mostafa J. EmpiricalEvaluation of Explicit Versus Implicit Acquisition of User Profiles in Information Filtering Systems[C]//Proceedings of the 4th ACM conference on Digital libraries. New York：ACM，1999：238-239.
14 Teevan J，Dumais S T，Horvitz E. PersonalizingSearch via Automated Analysis of Interests and Activities[C]// Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York：ACM，2005:449-456.
15,23 郭光明. 基于社交大数据的用户信用画像方法研究[D]. 合肥：中国科学技术大学,2017.
16 刘蓓琳，张琪. 基于购买决策过程的电子商务用户画像应用研究[J]. 商业经济研究，2017(24)：49-51.
18 王丹. 基于主题模型的用户画像提取算法研究[D]. 北京：北京工业大学,2016.
20,26,31 单晓红，张晓月，刘晓燕. 基于在线评论的用户画像研究——以携程酒店为例[J]. 情报理论与实践，2018，41(1)：99-104.
21 Carman M J，Baillie M，Crestani F. TagData and Personalized Information Retrieval[C]//Proceedings of the 2008 ACM Workshop on Search in Social Media. New York：ACM，2008：27-34.
22 Wetzker R，Zimmermann C，Bauckhage C，et al. ITag，You Tag：Translating Tags for Advanced User Models[C]//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York：ACM，2010：71-80.
24 李竟. 社会化问答社区协作性内容生产系统探究[D]. 南京：南京大学,2018.
25,30 张海涛,崔阳,王丹,等. 基于概念格的在线健康社区用户画像研究[J].情报学报,2018,37(9):912-922.
29 陈志明,胡震云.UGC网站用户画像研究[J].计算机系统应用，2017,26(1):24-30.
32 Zhang X，Zhao L，Zong L，et al. Multi-view Clustering via Multi-manifold Regularized Nonnegative Matrix Factorization[C]// 2014 IEEE International Conference on Data Mining (ICDM). IEEE Computer Society，2014.

（收稿日期：2019-06-10）

（上接第63页）

13 Kontostathis A,Galitsky L M,Pottenger W M,et al. A Survey of Emerging Trend in Textual Data Mining[M]. Survey of Text Mining:Clustering,Classification,and Retrieval,New York：Springer Verlag,2004：185-224.
14,16 刘自强,王效岳,白如江. 基于时间序列模型的研究热点分析预测方法研究[J]. 情报理论与实践,2016,39(5):27-33.
15 陈伟,林超然,李金秋,等. 基于LDA-HMM的专利技术主题演化趋势分析—以船用柴油机技术为例[J]. 情报学报,2018,37(7):732-741.
17 李海林,梁叶,王少春. 时间序列数据挖掘中的动态时间弯曲研究综述[J]. 控制与决策,2018,33(8):1345-1353.
18 张文秋,房磊,杨健,等. 基于Landsat时间序列的湖南省会同县杉木人工林干扰历史重建与林龄估算[J]. 生态学杂志，2018,37(11):3467-3479.
19 沈文娟,李明诗,黄成全. 长时间序列多源遥感数据的森林干扰监测算法研究进展[J]. 遥感学报,2018,22(6):1005-1022.
20 杨斌清,张希琳. 基于ARIMA时间序列模型的稀土氧化物价格预测研究[J]. 中国稀土学报,2017,35(5):680-686.
21 张美英,何杰. 时间序列预测模型研究综述[J].数学的实践与认识,2011,41(18):189-195.
22 Cavanaugh J E . Model Selection：Bayesian Information Criterion[M]// Wiley StatsRef：Statistics Reference Online. American Cancer Society，2016:1-3.
23 解素芳,王朋,焦淑静.基于信息构建的高校档案馆网站评价指标体系设计[J].档案学通讯,2010(6):53-56.
24 李萍萍,田原,刘慧,王朋.基于信息构建的医学院校图书馆网站评价指标[J].医学信息学杂志,2012,33(8):47-50.
25 伍玉伟.信息构建(IA)与信息组织的比较研究[J].图书馆论坛,2006,6(4):49-51.
26 刘伟,郝俊勤.信息组织与信息构建[J].情报资料工作,2009(1):27-29.
27 侯荣理.信息构建在网络信息资源整合中的应用[J].图书馆学研究,2006(11):59-60,88.
28 陈开慧,程小清.信息构建在网络信息资源中的实证研究[J].玉林师范学院学报(自然科学版),2008,29(3):124-127.
29 王晰巍,靖继鹏,范晓春.知识构建对知识管理的优化及实证研究[J].图书情报知识,2008(5):34-37,61.
30 余小鹏.基于信息构建的电子商务网站搜索系统研究[J].情报科学,2011,29(5):778-781.
31 陆怡洲.基于信息构建理论的图书馆网站标识系统研究[J].图书馆建设,2012(7):5-8.
32 周晓英.政府网站信息构建的特点:加拿大政府网站案例研究[J].情报理论与实践,2008,31(1):51-54.
33 王晓艳,胡昌平.基于用户体验的信息构建[J].情报科学,2006,24(8):1235-1238.

（收稿日期：2019-07-14）