计算机研究与发展 DOI:10.7544/issn1000-1239.2018.20180139
Journal of Computer Research and Development 55(9):1903-1919,2018

开放互联网中的学者画像技术综述

袁 莎[1] 唐 杰[1] 顾晓韬[2]
[1](清华大学计算机科学与技术系 北京 100084)
[2](伊利诺伊大学厄巴纳-香槟分校计算机科学系 伊利诺伊州厄巴纳-香槟 61801)
(yuansha@tsinghua.edu.cn)

A Survey on Scholar Profiling Techniques in the Open Internet

Yuan Sha[1], Tang Jie[1], and Gu Xiaotao[2]
[1](*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)
[2](*Computer Science Department, University of Illinois at Urbana-Champaign, Urbana-Champain, IL* 61801)

Abstract Scholar profiling from the open Internet has become a hot research topic in recent years. Its goal is to extract the attribute information of a scholar. Scholar profiling is a fundamental issue in large-scale expert databases for finding experts, evaluating academic influence, and so on. In the open Internet, scholar profiling faces new challenges, such as large amount of data, data noise and data redundancy. The traditional user profiling methods and algorithms cannot be directly used in the user profiling system in the open Internet environment. In this paper, the existing technologies are summarized and classified to provide reference for further research. Firstly, we analyze the problem of scholar profiling, and give a general overview of the information extraction method, which is the basic theory of user profiling. Then, the three basic tasks of scholar profiling including scholar information annotation, research interest mining and academic impact prediction are introduced in detail. What's more, the successful application system of scholar profiling called AMiner is introduced. Finally, open research issues are discussed and possible future research directions are prospected.

Key words user profiling; scholar profiling; information extraction method; research interest mining; academic impact prediction

摘 要 开放互联网中的学者画像工作是近年来的研究热点问题.学者画像的目标是提取学者各维度的属性信息进行信息挖掘和分析应用.学者画像技术是大型智库实现专家发现、学术影响力评估等功能的关键.在开放互联网中,学者画像面临数据量大、数据噪音和数据冗余等新挑战.这使得传统的用户画像理论、模型和方法无法直接无缝地移植到开放互联网环境下的用户画像系统中.针对这些挑战,对现有学者画像技术进行了总结和分类,为进一步的研究工作提供参考.首先分析了学者画像问题,对学者画像的基础理论——信息抽取方法——进行了总体概述,详细总结了各种可用模型与方法;对实现学者画

收稿日期:2018-02-26;**修回日期**:2018-06-25
基金项目:国家自然科学基金优秀青年科学基金项目(61222212);国家自然科学基金项目(61806111);国家"八六三"高技术研究发展计划基金项目(2015AA124102)
This work was supported by the National Natural Science Foundation of China for Excellent Young Scientists (61222212), the National Natural Science Foundation of China (61806111), and the National High Technology Research and Development Program of China (863 Program) (2015AA124102).
通信作者:唐杰(jietang@tsinghua.edu.cn)

像的基本任务包括学者信息标注、研究兴趣挖掘和学术影响力预测进行了详细阐述；介绍了学者画像应用实例 AMiner 系统；对未来重点的研究内容和发展方向进行了探讨和展望.

关键词 用户画像；学者画像；信息抽取方法；研究兴趣挖掘；学术影响力预测

中图法分类号 TP182

用户画像是指通过获取构成用户模型的不同维度属性信息(如人口统计学特征、兴趣偏好和行为模式等)进行信息挖掘和分析应用的过程. 在互联网时代，用户画像是实现精准化推荐和个性化服务的基石，在电子商务、社会网络分析以及互联网服务等众多领域有着广泛的应用. 例如在电子商务系统中，用户的历史购物习惯和偏好对商品的定向推荐和营销有着极其重要的作用；在社会网络中，用户的个人信息和社交交互数据能被用于好友推荐和社群发现；电信网络服务类行业依托用户属性实现个性化的订制服务.

虽然在不同的应用中，实现用户画像的具体参数有所不同，但是实现用户画像的基础技术是通用的. 为了表述的准确与清晰，以学者画像为例进行用户画像相关基础理论与技术要点的阐述与分析. 面向科研学者的用户画像技术为学术同行分类、专家推荐等功能提供了关键支持，科研学者画像问题已经受到了广泛的关注，以研究学者为中心的学术智库在国家自然科学基金委、科技部、中国工程院等权威部门展开应用.

传统的用户画像通常被当作是一个工程问题，构成用户模型的属性值是由人工收集的，或者是由用户主动提供的. 然而，人工数据收集往往需要花费大量时间和资源，对人工资源要求高的同时，数据获取的效率极其低下. 此外，用户通常不愿意花费时间和精力填写构建用户模型需要的属性信息，由用户输入的信息很多情况下是不完整的或者不一致的. 传统用户画像面临的这些问题使得建立大规模高质量的用户画像数据库成为难题.

近年以来，通过先进的计算机技术进行自动信息抽取逐渐取代手工方法成为主流. 这类方法首先收集大规模的电子文档，然后分别利用预定义的规则或者特定的机器学习模型抽取各项信息. 例如，使用交互式信息提取方法帮助用户将非结构化数据(如网页文档或电子邮件)输入数据库[1]，协助用户填写数据库字段，减少用户负担的同时保证输入数据的完整性；通过级联混合模型从简历中自动抽取结构化信息，实现了简历数据库的自动构建[2]. 然而，这类分别抽取不同属性信息的方法非常低效，原因有 2 点：1)对于每一个属性，必须定义一个特定的规则，或者通过监督学习训练一个特定的机器学习模型，属性的增多导致规则和模型的增多，大量不同的规则和模型非常难以维护；2)独立的各种规则或模型不能充分利用不同属性之间的依赖关系. 此外，在开放互联网中，这类方法虽然能自动抽取信息，然而却难以应对真实数据的动态变化. 仅以著名社交网络 Twitter 为例，其每日活跃用户量达到 2.5 亿以上，高峰期能够产生每秒 14 万条信息的数据量，通过线下数据库动态追踪、实时更新网络信息是非常困难的，难以保证数据的时效性[3].

开放互联网中的数据，尤其是万维网网页数据(Web 数据)以指数级的速度迅猛增长. 第 41 次《中国互联网络发展状况统计报告》指出，截至 2017 年 12 月，我国网民规模达 7.72 亿，移动互联网接入流量比上年同期累计增长 158.2%，呈现指数增长趋势[4]. 目前，基于 Web 的用户画像研究旨在从非结构化的 Web 网页文本中发现和挖掘结构化的用户信息. 例如，文献[5]基于 GATE 系统进行 Web 网页的分割和信息抽取；文献[6]提出一种无监督 Web 用户画像框架，在不依赖人工标注的情况下实现自动抽取. 然而，这些方法在当今网页数据的快速增长下仍然面临着大规模数据带来的存储和计算压力.

海量网络数据易于获取，并且蕴含着丰富的信息，这为大规模用户信息抽取提供了新的渠道和机遇，同时也面临 3 个特点和挑战：

1) 数据量大. CINIC 的统计数据表明：截至 2017 年 12 月，中国互联网中 Web 网页数达到 2 604 亿个. 即使在大型分布式系统的支持下，抓取、下载、索引这些网页数据需要耗费大量的存储和网络资源，传统的数据挖掘和信息抽取算法在如此巨大的搜索计算空间中面临着效率瓶颈，甚至无法有效运行. 在开放互联网环境下，面对大规模真实数据的动态变化，如何进行高效的信息抽取是亟需研究的问题.

2) 数据噪音. Web 数据中除了蕴含丰富的信息外，还混杂着大量的噪声数据，这些噪声数据会干扰用户画像的质量，这是伴随大数据量而来的必然问

题.例如,采用搜索引擎的查询结果进行信息抽取时,特定查询词的搜索结果往往包含了一些无关的词条,错误的信息抽取结果会影响抽取精度.数据噪音是开放互联网中信息抽取系统面临的主要精度瓶颈.

3)数据冗余.开放互联网中,数据存在着大量的冗余信息,这些冗余信息蕴含着隐含的关系模式.重要的信息在不同信息源中通常会重复出现,充分利用冗余信息之间的关联关系能充分挖掘更多的有用信息,同时帮助提升抽取信息的准确度.

虽然用户画像理论及相关技术已经得到了较为广泛的研究,但是在开放互联网中的用户画像具有独特性,其所面临的数据量大、数据噪音和数据冗余等新挑战致使传统的用户画像理论、模型和方法均无法直接无缝地移植到开放互联网中的用户画像系统.近年来,许多研究人员都致力于开放互联网中的用户画像研究,在理论、模型和方法等多个研究领域都进行了开拓性的探索,并提出了创造性的研究成果.本文正是以此为背景,对开放互联网中用户画像的现有研究成果进行回顾,对相关的研究思路进行溯源和比较,并给出了学者画像系统的实现案例.

1　问题描述

科学技术的发展带来了大量的学术数据,对于学术数据的挖掘越来越受到研究者的关注,很多学术系统都致力于学术信息挖掘的研究,如 Libra,Rexa,DBLife 等.学术信息挖掘的主要研究内容有各种学术数据的结构化组织,用元数据记录各种数据,如论文、研究者、会议等,学术信息的结构化组织中论文的结构化组织相对容易,技术也比较成熟,例如 Citeseer,DBLP 都提供论文的结构化数据,列出了论文的作者、题目、发表的会议、引用的参考文献等.研究学者也是学术信息的重要数据,是学术数据挖掘的重要研究方向,同时也是搭建学术社会网络的基石.

学者画像的例子如图 1 所示.学者画像的基本目标是为每个学者建立档案,包含学者的各种属性:基本信息(如名字、照片、工作单位、职位等)、联系信息(如电话、通信地址、Email 等)、教育经历(如毕业学校、所获学位的专业和时间等)、发表的论文以及研究兴趣.对于学者画像而言,有些画像信息(如基本信息、联系信息、教育经历)可以从其主页或者 Web 网页中获取,有些画像信息(如发表的论文)需要从在线数字图书馆(如 DBLP,ACM 等)整合得到,其他信息(如研究兴趣)需要从已收集的信息中挖掘分析得到.

学者画像的数据模式如图 2 所示.完成学者画像的数据标注需要从非结构化的数据中抽取目标信息,如地址、职位、所在机构、联系方式等不同类别的属性信息.经过统计分析发现,学者信息的各个属性之间有依赖关系,有的属性之间存在强依赖关系.举例来说,科研学者的名字可以帮助识别其照片,因为照片的命名往往是其本人的姓或名.在描述个人的教育经历时,比如科研学者获得了博士学位(PhD),那么获得博士学位的专业(PhDmajor),获得博士学位的日期(PhDdate)很可能出现在同一句话中,或者一个列表中.

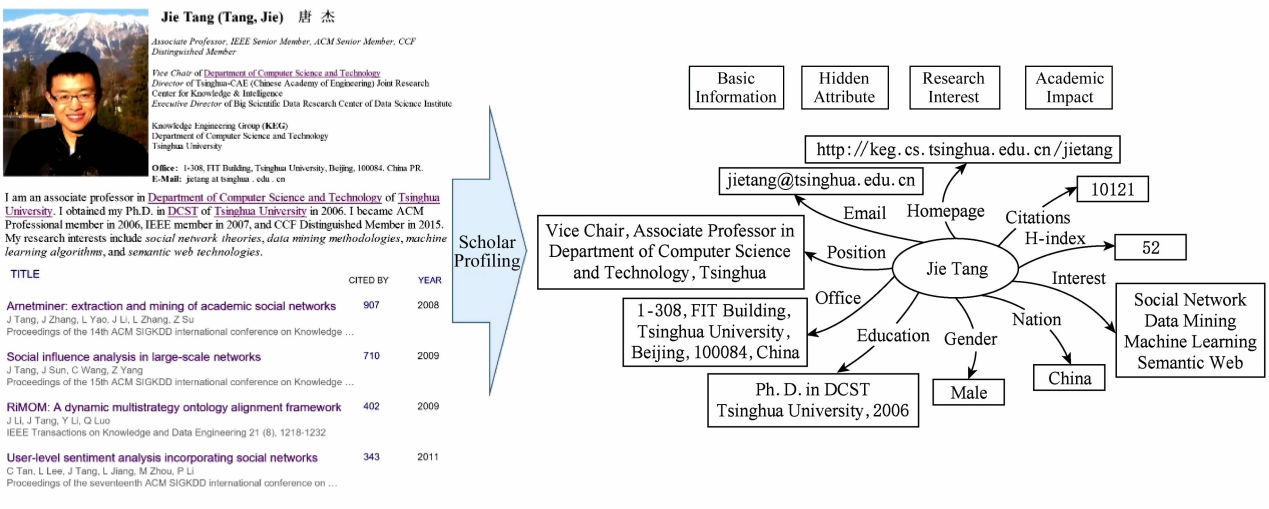


Fig. 1　An example of scholar profiling

图 1　学者画像示例

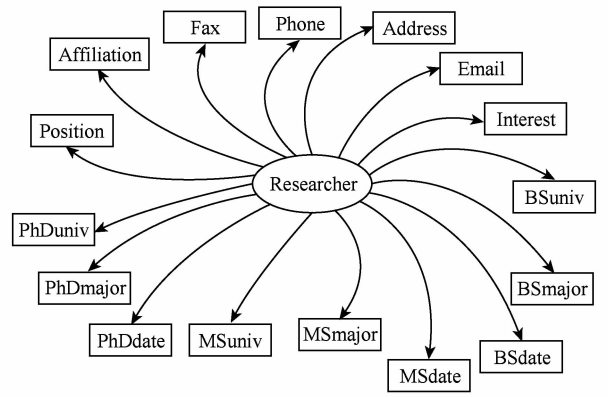


Fig. 2 Data pattern example of scholar profile

图 2 学者画像数据模式示例

由于 Web 数据本身的特性，针对 Web 用户的信息抽取任务需要解决 3 个问题：

1）快速信息检索. 对于不同类型的抽取任务构造合适的方法从 Web 中快速找到尽可能多的相关网页数据，从而避免遗漏有效信息.

2）排除数据噪音. Web 数据中除了丰富的有效信息外，同时混杂着许多噪音数据. 噪音数据提供了错误的信息，影响抽取精度. 在整合多源数据时，我们需要对噪音数据进行识别和筛除.

3）利用冗余数据. Web 数据中存在着大量冗余信息，其中又存在着隐含的关系模式. 例如，重要的信息会在不同信息源中重复出现. 利用冗余数据之间的关系能够捕捉到更多证据，帮助我们对所抽取信息的真伪进行更为准确的判断.

用户画像需要从非结构化数据中抽取目标信息，如地址、职位、所在机构、联系方式等，这往往依赖信息抽取方法及相关模型来实现. 信息抽取方法与模型是实现学者画像的理论基础，本文将在第 2 节详细介绍. 实现学者画像的 3 项基本任务将在第 3 节详细介绍，主要包括 3 个方面：

1）学者信息标注. 学者信息标注需要基于开源异构数据自动提取学者的相关描述信息，标注学者信息并建立学者个人档案，是实现学者画像的一项最基本的任务. 学者信息标注包括基本信息抽取和隐含属性预测，用户的隐含属性指难以从表层文本中直接抽取的属性数据，如性别、年龄等.

2）研究兴趣挖掘. 用户兴趣挖掘是指从用户数据中获取用户的偏好信息以及和用户相关的主题关键词，从而挖掘出用户兴趣. 研究兴趣挖掘主要应用于学者的研究方向发现，可用于学术合作推荐.

3）学术影响力预测. 论文被引数是评估学者学术影响力的重要指标之一，预测论文未来的被引数对学者科研水平评估及资助决策具有重要意义.

2 信息抽取方法

信息抽取方法是实现用户画像的基础理论，根据实现原理可将其划分为基于规则学习的方法、基于分类模型的方法和基于序列标注的方法.

2.1 基于规则学习的方法

基于规则学习的方法认为从大规模的自然语言数据中能够提炼和学习出频繁的规则或文本模式，并以此进行信息抽取. 例如“牛顿生于 1643.”这句例句中能够提取出“〈某人〉生于〈某年〉.”这样的模式. 将其与新的语料匹配时，即可从符合该模式的语料中抽取用户的生日信息. 依据具体实现方式的不同，基于规则学习的方法一般可划分为基于词典的方法和基于规则的方法.

2.1.1 基于词典的方法

早期传统的信息抽取系统采用基于词典的方法进行信息抽取，这类方法首先构建模式词典，然后利用词典从未标记的新文本中提取所需信息. 基于词典的方法实现的典型系统有 AutoSlog[7]，AutoSlog-TS[8] 和 CRYSTAL[9] 等. 实现这类系统的关键是如何习得模式词典，然后将其用于识别新文本中的相关信息.

AutoSlog 是第 1 个实现从文本训练集中学习模式词典的系统. AutoSlog 系统使用预定义的 13 种语法模式，例如下述句法“主语、直接宾语，或名词”是其中一种模式. AutoSlog 系统需要使用语法解析器生成句子的语法元素（例如主语、动词、介词短语），然后将生成的语法元素与给定的语法模式匹配，利用最佳匹配构建模式词典. AutoSlog 需要在提取模式前对文本进行标注，AutoSlog-TS 系统改进了这一缺点. AutoSlog-TS 不需要对输入数据进行完整的标注，只需要标注数据是否与主题相关. CRYSTAL 系统对少量已标记文本（称为种子词语）采用自助抽样法 Bootstrapping 从而生成词典. 具体地，CRYSTAL 基于 Bootstrapping 利用种子词典学习模式，然后使用已知模式标记更多同类的种子词语，这样就能持续增量地得到模式词典.

2.1.2 基于规则的方法

随后出现利用一般规则替代词典对文本进行信息抽取的系统，例如 $(LP)^2$[10]，DIPRE[11] 和 Snowball[12] 等. $(LP)^2$ 是经典的规则学习算法，基于 $(LP)^2$ 实现

的自动标注工具 Amilcare 能够从训练数据里自动学习规则.早期很多标注系统都是基于 Amilcare 实现的,如 S-CREAM[13],MnM[14] 和 Melita[15].这些系统能够在特定模板的网页上取得较好的标注效果,但是不能同时标注多种类型的信息.如果需要标注多种不同类型的信息,针对每一种新类型的信息,都需要重新学习一组规则,不能用于大规模的数据标注.

DIPRE 系统基于给定的已知事实例句,通过最长公共子句的方式归纳出规则,然后搜索与规则相关的更多例句,继续归纳和扩展规则.通过这种迭代式的半监督框架,DIPRE 系统能根据有限训练数据集自动搜索和扩展模式库,但是其严格的文本匹配方式会导致高遗漏率,并且迭代归纳过程中引入的错误种子数据会导致错误累积.Snowball 系统延续了 DIPRE 的半监督学习框架,但是放宽了模式匹配的条件,使得文本匹配的适用性大大提高.同时,其提出一系列方法以衡量学习到的模式与抽取到的目标信息的可信度,及时筛除数据噪音,从而减少迭代过程中的错误累积问题.后续工作 StatSnowball[16] 则引入了更多统计与机器学习的技巧,进一步提升效果.在数据驱动的模式学习基础上,PATTY[17] 引入文本解析树等语言学特征进行文本模式发现.

这些经典方法为文本模式学习提供了范式,然而也面临着标注数据不足、文本模式可扩展性不高的问题.近年来,随着 Freebase 等大规模知识图谱的发展,许多工作转而研究如何利用知识图谱进行远程监督学习.典型代表是 Riedel 等人在 2013 年提出的 Universal Schema 方法[18].他们借鉴协同过滤算法,将目标实体对看做用户,将实体关系看做商品,将信息抽取转化成商品推荐问题,即寻找目标实体对最有可能符合的实体关系.他们将文本模式与知识图谱中已有的实体关系共同作为学习目标,通过矩阵分解算法建立文本模式与目标实体的向量表达,从而更为广泛地衡量文本模式和目标实体间的隐含关系,以及文本模式间、文本模式与知识图谱关系间的隐含相似度,从而取得了很好的效果.

综合来说,基于规则学习的方法能够从大规模语料库中得到目标信息实体在文本中的频繁模式,并通过较为严格的文本匹配从目标文本中抽取信息,从而具有较高的抽取精度,但在查全率上表现较差,缺乏可扩展性,不适于大规模数据的信息抽取.

2.2　基于分类模型的方法

近年来,机器学习理论在信息抽取领域得到了成功的应用,监督机器学习方法为学者画像中的信息抽取任务提供了强有力的工具.基于分类模型的信息抽取方法将信息抽取转化为关系分类问题,即判断 2 个目标实体间是否满足目标关系,并给出判决结果.例如判断"牛顿"与"1643 年"之间是否存在"〈出生于〉"关系.依据信息抽取方法的发展趋势,基于分类模型的方法可分为基于机器学习的方法和基于深度学习的方法.

2.2.1　基于机器学习的方法

传统经典的分类模型有逻辑斯蒂回归(logistic regression, LR)、支持向量机(support vector machine, SVM)、决策树(decision tree, DT)和朴素贝叶斯(Naïve Bayes, NB)等.分类模型包含学习和预测 2 个阶段.在学习阶段,分类模型依据训练数据集训练模型;在预测阶段,训练出的模型被用于预测一个未标记的实例属于正例或负例.在二分类场景下,这些模型接受一个数据点的特征向量,并给出其属于正例或负例的预测.二分类模型可以通过对多个类别标签进行"属于/不属于"的二值判断扩展成多分类模型.

将这类模型用于解决文本信息抽取问题时,首先通过命名实体识别等预处理过程得到候选实体,然后从上下文文本中抽取特征,并通过分类模型预测该候选实体是否是正确的信息.特征的构造是影响分类模型识别准确率的重要因素.用于文本信息抽取的特征主要分为语义特征和语法特征.语义特征指句子各成分间的依赖关系,取决于目标实体在句子的依赖解析树中的依赖路径;语法特征指句子和实体上下文的浅层特征,常见的包括:目标实体间的语序、目标实体的词表示、目标实体的实体类型、目标实体的 POS 标签、整句句子的词袋表示、目标实体间的解析树路径、目标实体上下文的 n-gram 和 skip-gram 特征.

分类模型首先将这些特征处理成数值化的特征向量,然后利用有标签数据进行训练,对无标签的数据进行预测.其不足在于仍然依赖于人工定义的特征集合,使得其难以捕捉全部有用特征.同时,分类模型使用的函数簇也往往比较简单,难以建模复杂的非线性关系.

2.2.2　基于深度学习的方法

深度学习方法利用神经网络模型和词的分布式表达解决分类任务.与基于特征的分类模型相比,深度学习模型有两大优势.首先,深度学习模型中,单词通过 word2vec[19] 等词嵌入技术转化为词向量,包

含了更多语义层面的隐含信息,从而使得模型能够捕捉到词语层面的相似关系;更重要的是,深度学习模型能够进行表示学习,从原始数据中自动学习得到有用的特征,避免了人工特征构造,同时能够产生更为有效的特征表示.深度学习方法在信息抽取问题上主要应用卷积神经网络和递归神经网络.

卷积神经网络将句子中的单词以词向量的形式叠加成句子矩阵,通过卷积和池化的操作捕捉词与词之间的语义特征,并且能够应对句子中词语交换的问题.文献[20]在槽填充任务上比较了卷积神经网络与传统分类模型.其中,卷积神经网络将句子以目标实体为分隔拆分成3个部分,对每个部分分别进行卷积操作,从而得到上下文的特征表示,并对目标实体进行分类.结果显示,卷积神经网络的性能要高于传统分类模型.

递归神经网络为每个词学习一个隐层表示,同时将上一个词的隐层作为下一个词的输入,从而捕捉上下文关系.全局最后一个词的隐层中包含了全句的语义信息.文献[21]提出将长短时记忆机制加入递归神经网络模型中进行信息抽取,使得模型能够适时地忽略一些信息,从而取得更好的效果.文献[22]在此基础上做了改进.他们在预处理时首先通过递归神经网络得到全句的向量表示,并将其加入模型的初始输入,从而在学习单个词语时考虑全句信息,提高了信息抽取的精度.

较之基于规则学习的方法,分类学习模型能够自动学习特征与标签间的关系,避免了模式匹配带来的高遗漏率,从而在效果上有较大提升.然而,这些模型没有考虑目标实体之间的关系这一重要信息,同时难以像规则学习那样加入先验知识帮助求解.

2.3 基于序列标注的方法

信息抽取领域常用的序列标注模型有最大熵Markov模型(maximum entropy Markov mode, MEMM)[23]、条件随机场(conditional random fields, CRFs)[24]、动态条件随机场(dynamic conditional random fields, DCRFs)[25]和树状条件随机场(tree conditional random fields, TCRFs)[26]等.基于序列标注的方法一般基于条件概率模型.条件概率模型指给定观察值序列 X,找出其对应的状态序列 Y,使得 $P(Y|X)$最大.比较流行的条件概率模型包括最大熵模型(maximum entropy, ME)[27],由最大熵模型和序列模型结合引申出了MEMM和CRFs.

定义1. 最优状态序列.令 X 表示待标注数据的随机变量,$X=(x_1,x_2,\cdots,x_n)$,Y 表示标注结果的随机变量,$Y=(y_1,y_2,\cdots,y_n)$.在模型的训练阶段,根据观察到的数据以及数据的状态标签,最大化条件概率$P(Y|X)$,估计模型中各个特征对应的参数的值.在模型的测试阶段,寻找一个状态序列,使得条件概率最大,即 $Y^*=\max\limits_{Y}P(Y|X)$.

2.3.1 最大熵Markov模型MEMM

在最大熵Markov模型MEMM中,定义某个状态的概率为 $p(y_i|x)=p(y_i|y_{i-1},x)$.该公式表明,某个序列的第 i 个观察值取某个状态的条件概率仅仅与该序列中前一个状态的取值有关.图3画出了MEMM的图形结构.

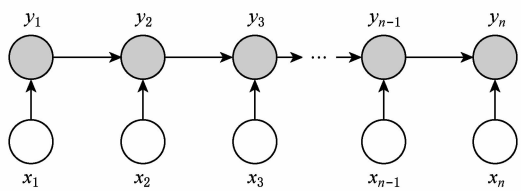


Fig. 3 The structure diagram of MEMM

图3 MEMM结构图

依据MEMM的定义,给定观察值序列 X,定义在该观察值上的状态值序列 Y 的条件概率为

$$p(Y\mid X)=p(y_1\mid x_1)\prod_{i=2}^{n}p(y_i\mid x_{i-1},y_{i-1}).$$

MEMM的模型定义为

$$p(y_i\mid y_{i-1},x)=\frac{1}{Z(y,x)}\exp\sum_k\lambda_k f_k(y_{i-1},x),$$

其中,$Z(y,x)$是归一化函数,它使得每个节点取所有状态的概率之和为1;λ_k 是待估计的参数;f_k 是特征函数.我们可以用通用迭代算法(generalized iterative scaling, GIS)算法或改进迭代尺度算法(improved iterative scaling, IIS)从训练集合学习参数的值[23].特征函数一般定义为谓词公式,也就是二值函数,或者叫布尔函数.在测试阶段,对于序列模型,通常用Viterbi解码这种动态规划方法求解最可能的状态序列.

2.3.2 条件随机场CRFs

MEMM对于序列中每个节点都归一化,这会导致label-bias问题[24].为了解决该问题,研究者们提出了条件随机场CRFs模型[24].

定义2. 条件随机场.令 $G=(V,E)$为无向图,其中 V 是节点集合,E 是边集合,$X=\{X_v|v\in V\}$表示 G 中所有节点的值,$Y=\{Y_v|v\in V\}$表示对 X 的标注结果.如果(X,Y)满足Markov性质[28]:

$$p(Y_v \mid X, Y_u, u \neq v, \{u,v\} \in V) \equiv p(Y_v \mid X, Y_u, \{u,v\} \in E),$$

即每个随机变量相对于 G 中所有变量的条件概率等价于它相对于 G 中所有相邻节点变量的条件概率,则称(X,Y)为条件随机场.

根据最大熵原理,可将 CRFs 中的条件概率定义为

$$p_\lambda(y \mid x) = \frac{1}{Z_\lambda(x)} \exp\left(\sum_{t=1}^{T} \sum_k \lambda_k f_k(y_{t-1}, y_t, x, t)\right),$$

其中:Z_λ 是归一化因子,计算为

$$Z_\lambda(x) = \sum_y \exp\left(\sum_{t=1}^{T} \sum_k \lambda_k f_k(y_{t-1}, y_t, x, t)\right),$$

其中,f_k 是以 y 和 x 为参数的特征函数,λ_k 是模型需要学习的参数(可以看作对应特征的权重).序列模型的学习,即是从训练样本中估计参数 λ_k 的值.用学习到的模型进行标注时,选择条件概率最大的标注序列 y^*,这与最大熵 Markov 模型类似,即:

$$y^* = \arg\max_y p(y \mid x).$$

很多方法都可以用来求解这个优化问题,传统的求解最大熵模型的算法如 IIS,GIS 都可以用来获得 CRF 模型的参数[29].共轭梯度法算法(conjugate gradient, CG)[30]和 L-BFGS 算法(limited-memory quasi-Newton)[31]能取得比较好的训练效果,投票感知器算法(voted perceptron, VP)算法[32]也能够有效地训练出 CRFs 模型的参数.由于 CRFs 的求解代价较大,目前仍然有很多研究者探讨求解 CRFs 的有效方法,如不采用 Maximum Likelihood 来求解 CRFs,而采用 Pseudo-Likelihood[33]求解 CRFs,对于复杂的 CRFs 模型,又有研究者利用 Piecewise 的方法[34],或者采用 Piecewise 结合 Pseudo-Likelihood 的方法求解复杂的 CRFs 模型.

CRFs 引入归一化标注因子,解决了 MEMM 中存在的 label-bias 问题.CRFs 克服了 MEMM 模型的缺点,它对整个序列做归一化,而不是对序列中的某一状态做归一化.用 CRFs 进行信息抽取实质上是将信息抽取问题转化为句子各部分的序列标注问题,即在观察到句子的各个实体部分情况下,推测它们对应的标签值.

2.3.3　其他条件概率模型及小结

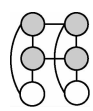
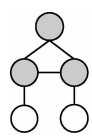
事实上像 CRFs 的定义一样,任何具有 Markov 性的图模型都可以称为条件随机场,目前有很多不同的条件随机场模型,它们不仅能模拟序列数据,而且能够模拟各种复杂结构的数据,如网格状数据,树状数据、甚至一般的图状数据.

动态的条件随机场 DCRFs 推广了线性的条件随机场.在 DCRFs 中,每一个片都是一个小型的贝叶斯网络,片与片之间的交互相当于线性 CRFs 的状态转移.树状条件随机场 Tree-CRF(TCRFs)可以描述信息之间的层次依赖关系.该模型是无环结构,求解起来相对容易.

表 1 列出了各种序列标注模型、图形表示以及求解算法.其中,MEMM 是最大熵框架下的序列模型;CRFs 最简单的形式是线性 CRFs;CRFs 还包含一般的图模型,例如 DCRFs 和 TCRFs 刻画了数据的多种依赖关系,能够更好地模拟数据,但求解更为复杂,可用贝叶斯网络里的 Belief Propagation 算法求解.

Table 1　Comparison of Conditional Probability Model

表 1　条件概率模型比较

Model	Structure Diagram	Algorithm
MEMM		IIS, GIS
CRFs		L-BFGS
DCRFs		L-BFGS Belief Propagation
TCRFs		Belief Propagation

Notes: The hollow circles represent the input, the solid circles represent the output.

较之规则学习和分类模型,基于序列标注的方法能够对实体间的关系进行建模,可以描述目标信息之间的依赖关系,有助于捕捉到更多信息,提高信息提取的准确性.

3　学者画像基本任务

信息抽取方法是实现学者画像的基础,有了基础理论的支持,本节对用户画像的 3 项基本任务:学者信息标注、研究兴趣挖掘以及未来影响力预测的相关工作进行概述.其中,学者信息标注分为基本信息标注和隐含属性预测,例如主页、邮箱、职位及办

公地址等信息为基本信息,此类基本信息可以从文本数据中显式抽取,性别和国籍则可能是需要预测的隐含属性;研究兴趣可以由学者主页提取,或者由学者发表的论文内容提取;学者的学术影响力通常由 H-index 值和论文引用数体现.

3.1 学者信息标注

研究学者的信息大多包含在学者的个人主页或者介绍性网页中,包含在其发表的论文中.信息来源的复杂性和信息格式的多样性使学者的个人信息标注成为学术信息挖掘的一个重大挑战.

根据标注方法的自动化程度,可以将信息的语义标注分为手工标注、半自动标注和自动标注.手工标注研究学者的基本信息非常繁琐,并且耗时耗力.已有研究工作表明自动标注能够从网页中提取有效信息,验证了自动标注的可行性和有效性.半自动标注利用一个预先制定的模板,或者针对每个属性学习出一个特定的模型来解决各个属性值的提取问题.但是,采用分别提取各个属性的技术效率很低.半自动标注技术先天上存在 2 个缺点:针对个人信息的每一个属性,都必须定义一个特定模板,或者学习一个特定模型,属性的增多导致模板和模型的增多,这些模板和模型比较难维护,训练时间也会很长;分散的规则或者模型不能够利用各个属性之间的依赖关系,而开放互联网中的 Web 数据特点是各个属性之间存在很强的依赖关系.自动标注模型需要利用属性间的相互依赖关系去提高识别各个属性的准确度.

定义 3. 学者信息自动标注模型.模型由 2 部分组成:学习和标注.学习模块的输入是已标注好的文档(即训练文档集$\{D_1,D_2,\cdots,D_n\}$和其对应的标注结果$\{y_1,y_2,\cdots,y_n\}$),学习模块通过分析训练样本中标注信息的分布,学习输出标注模型;标注模块的输入是待标注文档集(即测试文档集$\{D_{n+1},D_{n+2},\cdots,D_{n+m}\}$),标注系统利用标注模型识别测试文档中的语义信息,并利用本体进行描述,最后输出标注结果$\{y_{n+1},y_{n+2},\cdots,y_{n+m}\}$.

3.1.1 问题描述

学者信息标注指基于本体对包含个人信息的文档进行语义描述,生成本体的实例.图 4 给出了学者信息语义标注的实例.本体定义了:个人的名字(Name)、地址(Address)、联系电话(Phone)等.语义标注根据本体从文档中抽出相应信息,实例中的标注输出基于本体的语义内容.例如:在输出的标注结果中,"+8610-62788788-20"被标注为 Phone,也称为本体元数据电话号码(Phone)的标注实例.

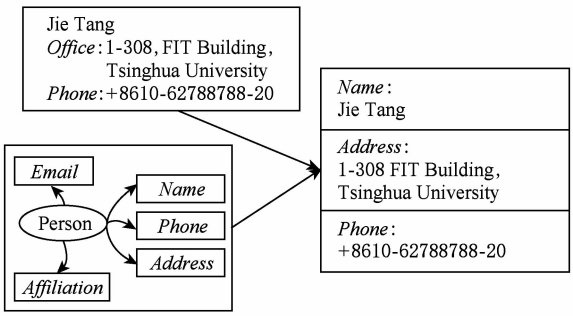


Fig. 4 An example of semantic annotation for scholars

图 4 学者信息语义标注示例

3.1.2 相关工作

早期研究工作集中在从特定结构的文档中抽取信息,例如借助交互式信息提取方法从电子邮件中抽取用户联系方式,协助用户构建数据库[1];Yu 等人通过级联混合模型从简历中自动抽取结构化信息[2],首先将简历分割成连续的区块并标注其类型,然后利用机器学习方法从特定的信息块中抽取特定信息,比如从联系信息块中抽取地址和电话等.这些研究集中于特定格式的文档,对源数据的格式有很大的限制,无法用于开放互联网环境中.当数据量增大时,这类方法面临着存储和计算的压力,可扩展性差.同时,这种被动抽取的方式只能从给定的数据中进行抽取,无法应对主动查询的要求,无法做到在线的信息抽取.

科研学者的信息大多包含在其个人主页或介绍性网页中,目前许多工作开始关注于从更广泛的 Web 数据中抽取用户信息.例如信息抽取系统 Artequakt[5]利用基于规则的抽取工具 GATE[35]从 Web 网页中抽取命名实体和它们之间的关系.文献[6]提出了从 Web 数据中以无监督的方式抽取信息.文献[36]提出了利用搜索引擎进行研究者画像的框架,首先通过分类模型从搜索引擎中找到用户的个人主页,然后从相对结构化的个人主页中抽取不同类别的文本信息.针对链式条件随机场无法建模层次化的 HTML 数据的问题,利用树形条件随机场模型 TCRF 进行网页内容的识别.通过将 HTML 文档转化成 DOM 树结构,利用 HTML 的层次化标签数据信息进行更精确的数据划分和预测,在电子邮件地址、机构信息等抽取问题上取得了 85%以上的 $F1$ 分数($F1$ Score),是目前相关问题的前沿方法.

除了个人基本信息外,也有一些工作研究用户隐含属性的抽取.用户隐含属性是指难以从表层文本中直接抽取的人口学信息,例如性别、年龄等.这些用户属性对用户行为和用户之间的交互有着深刻的影响,也是我们理解用户行为的关键信息,对于许多线上系统有着重要价值.然而,高质量的用户属性信息却非常难以获取,成为稀缺资源,也由此催生了许多尝试自动预测用户属性的研究工作.大多数相关工作在特定类型的用户数据领域开展,例如文献[37]尝试从用户的浏览行为中推测用户性别;文献[38]和文献[39]分别从游戏中的社交行为和搜索引擎的查询习惯2个角度出发预测用户性别.

在社会网络研究中,也有许多工作从年龄[40]、地理位置[41-42]、身份识别[43]等多个角度尝试利用社交网络信息预测用户属性.文献[44]细致地在移动网络中对用户属性的预测和观察,在大规模的语音电话、短信网络上,基于用户的通信行为利用概率因子图模型同时对用户的年龄和性别进行预测,同时提出了用户行为模式和隐含属性之间的关联.然而,他们的方法都是针对特殊数据高度订制的,扩展性较差,无法整合多源数据.一些工作意识到了多源数据的重要性,尝试通过整合多个数据源提高预测精度.例如文献[41]和文献[45-46]中的工作尝试整合Facebook,Google Plus以及Twitter的数据进行多源属性预测,取得了显著的精度提升.

基于名字的性别预测方法[47]具有较好的效果,该方法从Facebook中抽取大量用户信息,根据他们的姓名和性别生成词典,从而计算每个名字的用户从属于某个性别的概率.此方法简单有效,在实际实验中取得了近90%的精度,并被应用于Genderize等性别预测系统中.该方法的性能完全依赖于姓名-性别词典的质量和覆盖率,系统需要维护和扩充一个巨大的词典,却仍然难以保证高覆盖率.

目前针对Web数据的信息抽取工作中,层次化基本信息标注存在错误累积的缺陷.利用用户的行为数据进行隐含属性预测的研究工作,揭示了隐含属性与行为模式之间的相关关系,但在单向的隐含属性预测任务上依赖于用户属性与用户行为的关联度,难以保证预测精度.

3.2　研究兴趣挖掘

研究兴趣是学者画像的重要组成部分,其不仅是学者本身的研究心得或研究拓展方向的集中体现,也能从中窥视不同背景的学者对研究热点领域或学科研究趋势的关注度和敏感度.研究兴趣挖掘指从学者数据中获取学者研究的偏好信息以及和学者研究相关的主题关键词.

3.2.1　问题描述

研究表明仅有21.3%的学者会在其主页中给出研究兴趣[36],现有工作通常由学者本人发表的论文著作中提取其研究兴趣,采用概率话题模型求解.

定义4. 学者研究兴趣.论文z由一系列的词语w_i以及该词语在文中出现的概率$p(w_i|z)$来表示,也即$z=\{(w_1,p(w_1|z)),(w_2,p(w_2|z)),\cdots,(w_N,p(w_N|z))\}$,则研究学者$x$的研究兴趣可表示为$\{p(z|x)\}_x$.

3.2.2　相关工作

用户兴趣挖掘指从用户数据中获取用户的偏好信息以及和用户相关的主题关键词.许多早期的研究工作尝试从用户相关的文档中挖掘用户兴趣.例如,文献[48]收集用户对感兴趣网页的评价结果用于构建用户画像,依据用户画像信息推测用户对网站主题的偏好,由此使用搜索引擎可以快速获取用户感兴趣的特定主题页面.此外,他们发现当训练数据较少时,引入词典信息会增加用户兴趣挖掘的准确度,但是当数据量增大时提升效果不再明显.文献[49]开发了个性化的网页浏览器,能够自动学习用户信息,并用于帮助用户寻找感兴趣的网页;文献[50]研究了如何用启发式的方法获取用户兴趣关键词,将机器学习方法应用于用户画像.

近年来的研究工作更多地探索了用户行为数据与用户兴趣主题之间的关联.文献[51]提出了从Twitter数据中挖掘用户的兴趣关键词,基于上下文特征和行为特征,采用远程监督方法对政党候选人发布的内容进行建模,预测Twitter用户的政治偏好;文献[52]从用户行为轨迹数据出发挖掘用户偏好,将非平稳的、时间异构的用户轨迹在隐含随机环境中分解成短的随机步长,分解后的轨迹在短时间尺度上是平稳的,可使用Markov随机过程进行建模挖掘用户在不同时间的兴趣偏好;文献[53]通过一个统一的概率模型对用户行为背后的用户偏好以及他们的社会网络链接进行了研究,发现用户的兴趣与社会网络信息之间是互利的关系.

3.3　学术影响力预测

评估科学家过去和未来的潜在影响是人才招聘和资助决策的关键,论文引用数一直是评价学术影响力的重要指标.目前科技论文的数量飞速增长,预测学者已有论文的未来引用数对学者科研水平评估及资助决策具有重要意义.

3.3.1 问题描述

一篇论文在 d 时段$[0,T]$内获得的引用数为一个时间序列$\{n_d^t\}_{t=0}^T$,其中 n_d^t 表示论文 d 在时刻 t 获得的论文引用数.

定义 5. 论文引用数. 对给定的论文文献集合记为 D,$\mathrm{card}(D)=M$,一篇论文 $d\in D$ 在时刻 t 的引用数 n_d^t 定义为 $n_d^t=\mathrm{card}(citing_d^t)$,其中,$citing_d^t=\{\tilde{d}\in D,\tilde{d}\neq d:\tilde{d}^t \text{ cites } d\}$.

在论文引用数预测问题中,对任意论文 d 的输入为$\{(\boldsymbol{x}_d^0,n_d^0),(\boldsymbol{x}_d^1,n_d^1),\cdots,(\boldsymbol{x}_d^t,n_d^t),\cdots\}\in\mathbb{N}^K\times\mathbb{N}$,其中,$\boldsymbol{x}_d^t$ 是 K 维特征向量,$\boldsymbol{X}=\{\boldsymbol{x}_d^0,\boldsymbol{x}_d^1,\cdots,\boldsymbol{x}_d^t,\cdots\}$,$n_d^t$ 表示论文 d 在时刻 t 获得的实际论文引用数,一般地,$0=n_d^0\leqslant\cdots\leqslant n_d^t\leqslant\cdots\leqslant n_d^T=N_d$. 论文引用数预测的目标是学习一个预测函数 f,预测论文 d 在时刻 t 的引用数,即学习 $f(d\mid\boldsymbol{X},t)\rightarrow\hat{n}_d^t$,其中 $\hat{n}_d^t$ 是预测论文引用数.

3.3.2 相关工作

信息爆炸时代,随着科技文献数量的迅猛增长,只有很少部分文献获得广泛关注[54]. 用一种动态评估的方法预测单个项目的流行度的能力,在营销、政策制定和风险管理等领域都具有重要意义. 早期对未来流行度的预测可主要分为 2 类方法,每类都有已知的优势和局限性. 1)侧重于在项目集合上再现某些统计量[55],这类模型已经成功地理解了流行动态的基本机制,然而没有提供获取具体参数的方法,这些模型缺乏对个人行为的动态预测能力;2)将人气动态视为时间序列,通过时间相关性来进行预测流行度[56],尽管这类方法在某些领域取得了初步的成功,但是这些模型是确定性的,流行动态建模忽略了注意力的下降过程.

Yan 等人[57]引入引文数预测任务,基于科学出版物的内容、作者、地点和出版年设计特征. 为了获得作者排名,计算每个作者以前年份的平均引用数,并根据其他作者的数量确定排名;Yan 等人[58]扩大了特征空间,但结果仍然表明,作者排名是特征空间中影响最大的因素;Livne 等人[59]从 Microsoft Academic Search 中提取大量不同的数据集. 这个数据集包含 3 800 万篇论文,分为七大学术领域. 对于引文统计问题,他们根据作者的姓名、作者机构、地点、参考文献和论文内容构建特征. 通过使用 SVR 发现最重要的一组特征是基于引用网络的特征,即出版社和参考文献的影响因子是文献计量最显著的决定因素.

Shen 等人[60]提出了一个生成概率框架,基于加强泊松过程明确地建立了单个项目获得普及的过程,预测流行动态. 该模型结合了流行动力学的 3 个关键要素:表征内在吸引力的恰当参数、解释注意力老化效应的时间松弛函数,以及与流行动力学中"富者更富"效应对应的强化机制. 该模型的优点体现在 3 个方面:1)直接模拟个体关注的到达过程;2)渐进概率模型可以很容易地纳入贝叶斯框架来解释外部因素,从而提高预测能力;3)选择特定松弛函数的灵活性使其成为一个通用的框架,可以用来调整不同领域的人际动态.

Pobiedina 等人[61]依据图挖掘技术,将引文计数预测任务作为引文网络中链接预测问题,其中,论文的引用次数等于网络节点的入度,其出度对应于参考文献的数量,由于出度在过去几年中保持不变,因此出现新的链接意味着相应论文的引用次数增加. 在引用网络中利用频繁的图模式挖掘,基于挖掘模式计算新特征 GERscore,解决引文数预测问题.

4 学者画像系统实例①

AMiner[62]是一个学术科技大数据分析与挖掘系统. AMiner 自动从开放互联网中抽取学者信息,建立了 1.36 亿的学者档案及科技智库,为科研人员及机构提供学者搜索/推荐、专家发现、成果评价、技术发展趋势分析等知识服务及核心技术支持.

AMiner 系统的核心模型与算法包括:基于话题的影响力分析模型,自动生成实体之间基于不同话题层次的影响力强度;概率因子图模型用于识别网络中不同类型的关系(如师生关系,合作关系等);基于社会知识图谱的学者研究兴趣分析;学者多维度评价等核心算法. AMiner 系统的应用层提供了多种知识服务,包括:支持按权威度、地域、语种、性别等过滤条件的专家发现,按 H-Index、论文数、引用数、活跃度、社交性、领域多样性等学者成果多维评价,学者历年研究兴趣发展变化趋势分析,以及学者语义信息抽取、学者档案管理,权威机构搜索、话题发现与趋势分析、基于话题的社会影响力分析、即时社会关系图搜索、文献与审稿人推荐、学者的线上社交以及交互式文献阅读等多种功能及知识服务.

鉴于 AMiner 在学者画像领域的权威性,本节介绍 AMiner 系统里学者画像模块 3 个基本任务的实现机理.

① https://www.aminer.cn/

4.1　学者信息标注

AMiner 发布的研究成果显示约有 85.62%的研究学者来自于大学或科研机构,14.38%的研究学者来自公司.对于来自同一个公司的研究者,他们的网页可能共享相同的模板,基于规则学习的信息标注方法可能很有效.但是,不同的公司有不同的模板,很难定义一个统一的模板,有效地提取各种信息.对于来自学校的研究者,由于研究者各自的喜好不同,网页的布局和内容都千差万别.约 71.88%是个人主页,其余的网页是介绍性网页.这 2 种网页具有不同的特点,比如个人主页可能联系方式等信息以列表的方式给出,而介绍性网页多以自然语言给出研究者的各种信息,有可能缺失联系信息.此外,个人信息的各个属性的实例(各个属性相应的值)中,约 40%的属性以表格或列表形式给出,约 60%的属性则隐含在自然语言中.

在 AMiner 系统中,学者信息标注采用的是统一标注模型,包括 3 个主要步骤:主页查找,预处理和信息标注.在主页查找中,给定研究学者的名字,通过搜索引擎得到一系列网页,而后训练一个分类器来判定这些网页是否是个人主页或者包含很多研究者信息的介绍性网页,最后把确认的网页的 URL 作为个人信息的属性 *Homepage* 的值.

AMiner 采用条件随机场作为标注模型.条件随机场模型 CRFs 的目标函数为

$$p_\lambda(y \mid x) = \frac{1}{Z_\lambda(x)}\exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(y_{t-1}, y_t, x, t)\right),$$

其中,x 代表观察值,即网页中的 token,y 代表观察值相应的标签,即本文定义的研究者个人信息的各种属性.f 代表数据的特征,λ 是各特征的系数,需要通过训练数据学到.

进一步地,AMiner 引入先验知识进一步提高抽取精度.基于 Markov 逻辑因子图模型,通过关系因子建模候选实体间的冗余关系,利用一阶逻辑知识库引入先验知识,筛除不符合常识的错误实体,从而提高抽取系统的准确性.AMiner 设计了 3 种关系因子:局部属性因子、局部逻辑因子和关系逻辑因子.

局部属性因子定义了特征向量 x_i 和标签值 y_i 间的数值关系,局部属性因子表达式为

$$\phi_f(y_i \mid x_i) = \frac{1}{Z_f}\exp\{\sum_{f_l \in F} w_l f_l(x_i, y_i)\},$$

其中,F 为所有特征函数的集合,Z_f 将函数值归一化成概率值,即给定 x_i 的情况下对应标签值为 y_i 的概率.局部逻辑因子由一阶逻辑知识库中仅涉及单实体的局部逻辑给出,记第 m 条局部逻辑的示性函数为 u_m,当 x_i 和 y_i 满足该逻辑时 u_m 的值为 1,否则为 0,类比局部属性因子,局部逻辑因子的表达式为

$$\phi_u(x_i, y_i) = \frac{1}{Z_u}\exp\{\sum_{u_m \in U}\alpha_m u_m(x_i, y_i)\}.$$

关系逻辑因子由一阶逻辑知识库中描述实体之间关系的关系逻辑给出,记第 k 条关系逻辑的示性函数为 r_k,当 y_i 和 y_j 满足该逻辑时 r_k 的值为 1,否则为 0,类比局部逻辑因子,关系逻辑因子的表达式为

$$\phi_r(y_i, y_j) = \frac{1}{Z_r}\exp\{\sum_{r_k \in R}\beta_k r_k(y_i, y_j)\}.$$

引入 Markov 假设,即图模型中的任一变量独立于所有非邻节点的其他变量.在给定包含所有变量节点和因子节点的图模型 G 时,标签值 Y 的条件概率为

$$P(Y \mid G) = \frac{1}{Z}\exp\{\sum_{x_i \in X}\sum_{f_l \in F}\omega_l f_l(x_i, y_i) + \sum_{x_i \in X}\sum_{u_m \in U}\alpha_m u_m(x_i, y_i) + \sum_{x_i \in X}\sum_{r_k \in R}\beta_k r_k(y_i, y_j).$$

通过 Markov 逻辑因子图模型,得以在基于分类模型的信息抽取方法基础上加入先验知识和冗余实体间的关系,提高信息抽取精度.

相对于各类单独的方法,统一标注模型有 2 个优势:1)对于个人信息的各个不同属性的标注是相互关联的,而不是独立的.基于规则学习的方法和基于分类模型的方法都要针对每个属性单独学习规则或训练模型,它们不能同时标注各个属性.而统一模型能够克服这一缺点,在统一框架下同时标注个人信息的各个属性,并且达到更高的标注精度.2)个人信息有很多特定的属性和属性值,如果利用基于规则学习的方法或者基于分类模型的方法,我们必须学习特定的规则和分类器去处理各种不同的情况,这会导致模型个数增多,难以管理、控制.大规模标注采用这些方法是不可行的.相反,统一模型能够训练一个模型去标注不同类型的属性,一次就能够解决所有的标注子任务.

4.2　学者兴趣挖掘

与其他用户信息不同,兴趣关键词很难区分正确与否,多个关键词保留了用户不同角度的特征.学者研究兴趣没有准确的评判标准,多数情况依赖于人工标注.AMiner 系统采用主题模型对抽取到的兴趣关键词进行聚类,以期找出用户的兴趣主题.

AMiner 采用文档主题生成模型(latent Dirichlet allocation, LDA)作为生成模型. LDA 是一个用于描述离散数据的生成模型[63]. LDA 假定文本中的每个单词是由一些混合的话题产生的,每个话题都有一定的权重,即:

$$p(w)=\sum_{z}p(w\mid z)p(z),$$

而 $p(z)$ 又由一个 Dirichlet 分布产生. LDA 是潜在语义分析(latent semantic analysis, LSA)[64]和概率潜在语义分析(probabilistic latent semantic analysis, PLSA)[65]的延伸. 这些模型都是为了分析文本的潜在结构,比如话题. 话题能够捕获文本的语义信息,帮助找到文本里意思相同或相近的单词,这使得信息检索时能够找到和查询中的单词语义相关的文档而不只是包含查询单词的文档.

LDA 的贝叶斯网络结构如图 5 所示,图 5 中 K 为主题个数,M 为文档总数,V 是给定某个文档中的单词总数,α 和 β 分别是每个 Topic 下词的多项式分布和每个文档下 Topic 的多项分布的 Dirichlet 先验参数. LDA 模型中的变量总结于表 2.

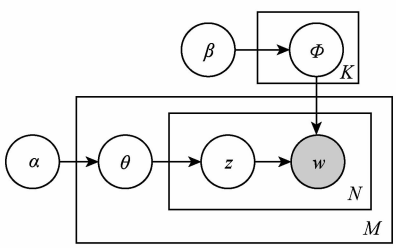


Fig. 5 An example of semantic annotation for scholars

图 5 LDA 结构图

Table 2 Notations in LDA

表 2 LDA 模型中的变量和标记

Symbol	Description
M	Number of documents in the corpus
K	Number of latent topics
V	Number of unique words in vocabulary
α	Hyper-parameter on the topic
β	Hyper-parameter on the word
θ	Notation for document-topic distribution $p(z\mid d)$, the size is MK
Φ	Notation for word-topic distribution $p(w\mid z)$, the size is KV
N	Number of words in one document
z	Topic assigned for each word in each document
w	Word in each document

根据 LDA 的图模型,给出 LDA 中的联合概率分布:

$$p(w,z\mid\alpha,\beta)=p(w\mid z,\beta)p(z\mid\alpha)=$$
$$\int p(w\mid f,z)p(f\mid\beta)\mathrm{d}f\int p(z\mid\theta)p(\theta\mid\alpha)\mathrm{d}\theta=$$
$$\left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^{W}}\right)^{K}\prod_{z=1}^{K}\frac{\prod_{w}\Gamma(n_{zw}+\beta)}{\Gamma(n_{z\times}+W\beta)}\times$$
$$\left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^{K}}\right)^{M}\prod_{d=1}^{M}\frac{\prod_{z}\Gamma(n_{dz}+\alpha)}{\Gamma(n_{d\times}+K\alpha)}.$$

LDA 模型中有一组隐含变量 z,要估计的参数是 θ 和 Φ. LDA 模型有多种求解方法,常用的有变分期望最大化算法(variational expectation-maximization, VEM)[63]、吉布斯采样(Gibbs sampling)[66]和期望传播方法(expectation propagation)[67]. 由于 Gibbs sampling 方法比较容易实现,AMiner 系统中采用了它. Gibbs sampling 使用采样的方法,构建 Markov 链,逼近目标概率分布. Gibbs sampling 需要求解条件概率 $p(z_i\mid z_{-i},w)$,从这个概率可以为当前单词抽样出一个 topic. 利用联合概率公式,可以得到条件概率的计算公式:

$$P(z=k\mid z,w)\propto\frac{n_{w,k,-i}+\beta}{n_{\cdot,k,-i}+W\beta}\frac{n_{d,k,-i}+\alpha}{n_{d,\cdot,-i}+K\alpha},\qquad(1)$$

其中,$n_{w,k,-i}$ 表示单词 w 被指派到 topic k 的次数,不包含当前这一次. "·"表示对这列求和,那么 $n_{\cdot,k,-i}$ 表示语料库中被指派到 topic k 的所有类别的个数,不包含当前这一次. 类似地可以理解 $n_{d,k,-i}$ 和 $n_{d,\cdot,-i}$. 从式(1),直观上可以看出第 1 项表示 $p(w\mid z)$,第 2 项表示 $p(z\mid d)$. 事实上,得到每个单词的 topic 之后,这些信息可以用于计算计算模型中的参数 θ 和 Φ:

$$\hat{\theta}_{dz}=\frac{n_{d,z}+\alpha}{n_{d,\cdot}+K\alpha},$$

$$\hat{\phi}_{zw}=\frac{n_{w,z}+\beta}{n_{\cdot,z}+W\beta},$$

有了这些参数后,我们可以用于计算:

$$p(w\mid d)=\sum_{z}p(w\mid z)p(z\mid d)=\sum_{z}\phi_{zw}\theta_{dz},$$

可以把查询拆成一个个单词,$p(q\mid d)$ 就可以表示成这些 $p(w\mid d)$ 的乘积[68].

4.3 论文引用数预测

AMiner 中的论文引用数预测模型同时考虑了 4 个现象:1)内在质量,即论文本身的质量;2)老化效应,即论文的新颖度会随着时间的推移而降低;3)马太效应,即"富者越富"现象;4)近因效应,最新研究成果容易获得更多关注. 基于这 4 个方面的考

虑,AMiner 采用循环神经网络(recurrent neural network, RNN)[69]和长短时记忆单元(long short term memory, LSTM)[70].

图 6 给出了 AMiner 中的论文引用数预测模型,采用 2 层 RNN,其中的神经元单元采用 LSTM. 给定一个时间序列$\{n_d^t\}_{t=0}^T$,RNN 能处理最近的时间事件,LSTM 单元具有记忆功能,能够处理和预测时间序列中间隔和延时较长的事件.

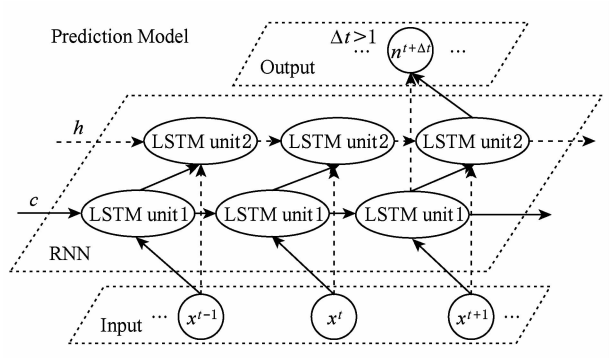


Fig. 6　Diagram of the citation count prediction model

图 6　论文引用数预测模型框架

论文的内在质量由$\{(x^0,n^0),(x^1,n^1),\cdots,(x^t,n^t),\cdots\}$体现. 老化效应由 LSTM 中的遗忘门模拟,遗忘门的公式为 $\Gamma_f^t=\sigma(W_f[h^{t-1},x^t]+b_f)$,其中 W_f 是遗忘门的权重,σ 表示 Sigmoid 函数. $\Gamma_f^t\in[0,1]$,0 意味着 LSTM 单元会清除指定信息,1 意味着信息会保留. LSTM 单元具有更新门,对应的公式为 $\Gamma_u^t=\sigma(W_u[h^{t-1},x^t]+b_u)$,$\Gamma_u^t\in[0,1]$. 当前信息为 $\tilde{c}^t=\tanh(W_c[h^{t-1},x^t]+b_c)$. 马太效应可由长期信息 $c^t=\Gamma_f^t c^{t-1}+\Gamma_u^t\tilde{c}^t$ 体现.

LSTM 单元的公式化如图 7 所示,其中 LSTM 的长期记忆单元为 c^t,当前信息记忆由短时记忆单元存储 $h^t=\Gamma_r^t\times\tanh(c^t)$. 当前工作单元的信息读取速度高于长期存储单元,模拟了近因效应. 最终,论文 d 在时刻 t 的引用数 $f(d|\boldsymbol{X},t)=\mathrm{softmax}(h^t)$ 给出.

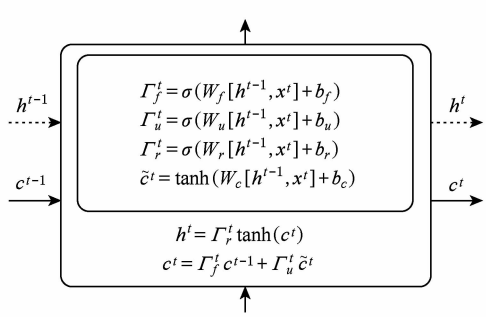


Fig. 7　Formulation of the LSTM unit

图 7　LSTM 单元的公式化

5　未来研究方向与挑战

开放互联网带来的海量数据为研究者画像问题带来了新的机遇,同时也带来了数据噪音、数据冗余等问题. 开放互联网中的学者画像研究取得了一定的进展,目前仍然是一个充满挑战和机遇的新兴研究领域.

在开放互联网中,学者画像信息抽取方法主要面临 3 方面的技术挑战:

1) 亿级网页抓取的工程实现. 针对开放互联网中的海量学术资源网页和链接,需要设计高效网页读写技术和链接抽取技术;针对亿级网页设计并行化爬取方法,需要实现高效的网页抓取和网页入库,网页前后链接发现和属性入库,相关的网页及链接库读写支持数百万级每秒查询率(query per second, QPS).

2) 智能学术实体与关系抽取. 实现基于监督学习和半监督学习结合的学术实体抽取技术;设计基于规则和先验知识的特定关系抽取方法,包括基于核函数、逻辑回归句法解析增强等机器学习手段;实现基于递归神经网络和卷积神经网络的开放学术关系抽取技术.

3) 基于关键特征的实体聚合. 利用学术成果中作者名、电子邮箱、单位、研究领域关键词等关键特征,基于概率模型的学术成果聚合技术;利用合作者网络,基于复杂网络技术的学术作者命名消歧技术,基于深度神经网络自动学习与实体相关的隐藏特征,学术成果特征自适应,提升聚合的准确度.

构建高精度学者画像库主要面临 4 个方面挑战:

1) 高精准学者画像库的构建方法. 需要基于多源异构科研行为数据构建出多维、立体的学术画像模型,完善科研人员与科研专家的画像库,此外,学者画像数据库的构建需要考虑基本属性维度、文献维度、时间维度、机构维度、地理位置维度、事件维度、主题维度、兴趣维度、行为维度、群体属性维度、心理维度等维度.

2) 多维度标签化技术. 需要解决科研人员、科技专家及科研行为关系的标签化方法,精确刻画科研人员、科技专家与科研行为的静态属性特征、动态行为特征、科研社群特征,此外,需要解决标签度量计算,引入画像可信度量打分机制,通过统计、排比、相似度计算等方法,构建〈用户,标签,可信度〉三元组,提高标签刻画精度. 自动化的高效标签化算

法也面临着实现大规模科研人员及行为高效画像的挑战.

3）科研学者画像样本验证数据集.用户画像数据主要是通过计算机采集数据进行计算推演的方式获取，缺乏一个准确的结果判定标准，需要研究测试判定样本的构造模型和机制，制作标准的测试样本集构造模型以及数据采集策略，所获取的测试样本应该具有典型的代表性和广泛性，形成科学、客观的科研学者画像样本验证数据集，用于客观评价学者画像的准确程度.

4）画像的增量更新与溯源技术.针对科研学者画像数据的增量更新频率，需要构建高效的触发器机制与传播更新机制，提升画像的时效性.针对科研学者画像中数据溯源模型的构建方法，设计数据溯源机制，通过对经典的数据溯源模型进行分析研究，需要制定适合科研行为画像溯源的模型及方法.

开放互联网中的学者画像研究取得了一定的进展，目前仍然是一个充满挑战和机遇的新兴研究领域，可以进行开拓式创新或继承式研究并取得成果的方向有很多，主要存在于4个方面：

1）面向多源信息的中文知识图谱实体与关系抽取.基于面向多源信息的学术知识图谱实体、属性和关系抽取技术，建立一个科学完整的科研行为命名实体分类体系，一方面用于指导算法研究，另一方面便于对抽取得到的实体数据进行管理.在此基础上，基于深度语义模型和半监督学习算法从相关语料中提取出科研行为实体之间的关联关系.可以考虑利用自然语言处理领域的深度神经网络语言模型、句法分析方法、篇章分析方法以及语言的可计算性理论等工具，特别是基于深度语义模型来获取数据源中实体的潜在语义表示，以及实体间关系的潜在语义空间，抽取复杂关系.

2）面向多源信息的科研行为实体对齐[71]与多尺度融合.基于半监督机器学习的自适应选择局部集体对齐和全局集体对齐的技术，基于概率模型（如贝叶斯网络、LDA模型、Markov逻辑网等）来学习实体间属性和结构的相似性去提高实体对齐的准确率和召回率，解决多源数据中科研行为实体命名规则、定义粒度、判别能力不同导致的数据质量问题.基于深度语义模型的科研行为实体融合技术，以直接优化消歧任务为训练目标，自动学习上下文和实体的特征表示和“上下文-实体定义”相似度度量，对上下文多个实体同时消歧，实现多尺度知识融合.

3）学术知识图谱关系扩展与推理.基于深度表示学习方法的知识图谱三元组编码技术，将它们的语义信息映射到低维的潜层特征表示空间（语法、语义空间），以推测知识图谱中存在的隐式知识.同时，基于一阶谓词逻辑为基础的符号逻辑知识表示方法，基于W3C标准知识描述系统，从大规模、半结构化或非结构化的数据源自动提取科研行为概念及其上下文关系，将符号逻辑模型中的推理机制应用于表示学习中，不断扩充和优化关系推理技术，提升大数据环境下科研行为知识表示学习的能力[72].

4）时、空多尺度场景下的知识图谱[73]主题演化与更新技术.根据实时采集的多源科研行为数据的动态变化，实时更新对应的学术知识图谱内容及网络结构，实现学术知识图谱中各类主题信息的实时更新.主要通过知识库语义模型得到实体和关系在知识图谱空间的潜层特征表示，并基于多模态的深度神经网络模型框架，实现多源异构学术数据的共享语义分析和动态更新.

6 总　　结

本文对学者画像的相关概念及方法深入研究的基础上，总结了实现学者画像的基本方法——信息抽取方法，以及3个基本任务包括学者信息抽取、研究兴趣挖掘以及学术影响力预测，给出了学者画像系统实现的实例分析.随着开放互联网规模的不断增长，开放互联网中的学者画像研究将会面临更多的问题和挑战.本文最后探讨了学者画像研究中值得探索的方向，供相关学者参考.

参 考 文 献

[1] Kristjansson T, Culotta A, Viola P, et al. Interactive information extraction with constrained conditional random fields [C] //Proc of the 19th AAAI Conf on Artificial Intelligence. Menlo Park: AAAI, 2004: 412-418

[2] Yu Kun, Guan Gang, Zhou Ming. Resume information extraction with cascaded hybrid model [C] //Proc of the 43rd Annual Meeting on Association for Computational Linguistics. Stroudsburg: ACL, 2005: 499-506

[3] Gu Xiaotao. Researcher profiling in the open Internet [D]. Beijing: Tsinghua University, 2017 (in Chinese)
（顾晓韬. 开放互联网中的研究者画像[D]. 北京：清华大学，2017）

[4] China Internet Network Information Center. The 41st China statistical report on Internet development [R/OL]. Beijing: China Internet Network Information Center, 2018 [2018-02-01]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201803/P020180305409870339136.pdf (in Chinese)

(中国互联网络信息中心. 第41次中国互联网络发展状况统计报告[R/OL]. 北京:中国互联网络信息中心,2018[2018-02-01]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201803/P020180305409870339136.pdf)

[5] Alani H, Kim S, Millard D E, et al. Automatic ontology-based knowledge extraction from Web documents[J]. IEEE Intelligent Systems, 2003, 18(1): 14-21

[6] Michelson M, Knoblock C. Unsupervised information extraction from unstructured, ungrammatical data sources on the World Wide Web[J]. International Journal on Document Analysis and Recognition, 2007, 10(3): 211-226

[7] Riloff E M. Automatically constructing a dictionary for information extraction tasks[C]//Proc of the 11th National Conf on Artificial Intelligence. Menlo Park: AAAI, 1993: 811-816

[8] Riloff E. Automatically generating extraction patterns from untagged text[C]//Proc of the 13th National Conf on Artificial Intelligence. Menlo Park: AAAI, 1996: 1044-1049

[9] Soderland S, Fisher D, Aseltine J, et al. CRYSTAL inducing a conceptual dictionary[C]//Proc of the Int Joint Conf on Artificial Intelligence. San Francisco: Morgan Kaufmann, 1995: 1314-1319

[10] Ciravegna F. an Adaptive algorithm for information extraction from Web-related texts[C]//Proc of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining. San Francisco: Morgan Kaufmann, 2001

[11] Brin S. Extracting patterns and relations from the World Wide Web[C]// Proc of the 1st Int Workshop on the World Wide Web and Databases. Berlin: Springer, 1998: 172-183

[12] Agichtein E, Gravano L. Snowball: Extracting relations from large plain-text collections[C]//Proc of the 5th ACM Conf on Digital Libraries. New York: ACM, 2000: 85-94

[13] Handschuh S, Staab S, Ciravegna F. S-CREAM: Semi-automatic creation of metadata[C]// Proc of EKAW'2002. Berlin: Springer, 2002: 358-372

[14] Vargas-Vera M, Motta E, Domingue J, et al. MnM: Ontology driven semiautomatic and automatic support for semantic Markup[C]//Proc of EKAW'2002. Berlin: Springer, 2002

[15] Ciravegna F, Dingli A, Iria J, et al. Multi-strategy definition of annotation services in melita[C]//Proc of ISWC'03 Workshop on Human Language Technology for the Semantic Web and Web Services. Berlin: Springer, 2003: 97-107

[16] Zhu Jun, Nie Zaiqing, Liu Xiaojiang, et al. StatSnowball: A statistical approach to extracting entity relationships[C]// Proc of the 18th Int Conf on World Wide Web. New York: ACM, 2009: 101-110

[17] Nakashole N, Weikum G, Suchanek F. PATTY: A taxonomy of relational patterns with semantic types[C]// Proc of the Joint Conf on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg: ACL, 2012: 1135-1145

[18] Riedel S, Yao Limin, Mccallum A, et al. Relation extraction with matrix factorization and universal schemas[C]//Proc of the 15th Annual Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2013: 74-84

[19] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint, arXiv: 1301.3781, 2013: 1-12

[20] Adel H, Roth B, Schütze H. Comparing convolutional neural networks to traditional models for slot filling[C]// Proc of the NAACL. Stroudsburg: ACL, 2016

[21] Yao Kaisheng, Peng Baolin, Zhang Yu, et al. Spoken language understanding using long short-term memory neural networks[C]//Proc of the 5th Spoken Language Technology Workshop. Piscataway, NJ: IEEE, 2015: 189-194

[22] Kurata G, Xiang Bing, Zhou Bowen, et al. Leveraging sentence-level information with encoder LSTM for semantic slot filling[C]//Proc of the 16th Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2016: 2077-2083

[23] Mccallum A, Freitag D, Pereira F C N. Maximum entropy markov models for information extraction and segmentation[C]//Proc of the 17th Int Conf on Machine Learning. New York: ACM, 2000: 591-598

[24] Lafferty J D, Mccallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proc of the 18th Int Conf on Machine Learning. New York: ACM, 2001: 282-289

[25] Sutton C, Rohanimanesh K, McCallum A. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data[J]. Journal of Machine Learning Research, 2007, 8(3): 693-723

[26] Tang Jie, Hong Mingcai, Li Juanzi, et al. Tree-structured conditional random fields for semantic annotation[C]//Proc of ISWC2006. Berlin: Springer, 2006: 640-653

[27] Berger A L, Pietra S A D, Pietra V J D. A maximum entropy approach to natural language processing[J]. Computational Linguistics, 1996, 22(1): 39-71

[28] Chellappa R, Jain A. Markov random fields[J]. Theory and Application, 1993 (1): 242-261

[29] Darroch J N. Generalized iterative scaling for log-linear models[J]. Annals of Mathematical Statistics, 1972, 43(5): 1470-1480

[30] Shewchuk J. An introduction to the conjugate gradient method without the agonizing pain[J]. Journal of Comparative Physiology, 1994, 186(3): 219-235

[31] Fei S, Pereira F. Shallow parsing with conditional random fields[C]//Proc of the 5th Conf of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Stroudsburg: ACL, 2003: 134-141

[32] Collins M. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms [C] //Proc of the 4th Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2002: 1-8

[33] Parise S, Welling M. Learning in Markov random fields: An empirical study [J]. Joint Statistical Meeting, 2005, 4: 1-7

[34] Sutton C, Mccallum A. Piecewise training for undirected models [C] //Proc of the 21st Conf on Uncertainty in Artificial Intelligence. Corvallis, Oregon: AUAI, 2005: 568-575

[35] Cunningham H, Maynard D, Bontcheva K, et al. GATE: A framework and graphical development environment for robust NLP tools and applications [C] //Proc of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2002: 168-175

[36] Tang Jie, Yao Limin, Zhang Duo, et al. A combination approach to Web user profiling [J]. ACM Trans on Knowledge Discovery from Data, 2010, 5(1): 1-38

[37] Hu Jian, Zeng Huajun, Li Hua, et al. Demographic prediction based on user's browsing behavior [C] //Proc of the 16th Int Conf on World Wide Web. New York: ACM, 2007: 151-160

[38] Szell M, Thurner S. How women organize social networks different from men [J]. Scientific Reports, 2013, 3(1): 1214:1-1214:6

[39] Bi Bin, Shokouhi M, Kosinski M, et al. Inferring the demographics of search users: Social data meets search queries [C] //Proc of the 22nd Int Conf on World Wide Web. New York: ACM, 2013: 131-140

[40] Sarraute C, Brea J, Burroni J, et al. Inference of demographic attributes based on mobile phone usage patterns and social network topology [J]. Social Network Analysis and Mining, 2015, 5(1): 39:1-39:18

[41] Li Rui, Wang Shengjie, Deng Hongbo, et al. Towards social user profiling: Unified and discriminative influence model for inferring home locations [C] // Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2012: 1023-1031

[42] Efstathiades H, Antoniades D, Pallis G, et al. Users key locations in online social networks: Identification and applications [J]. Social Network Analysis and Mining, 2016, 6(1): 66:1-66:17

[43] Joseph K, Wei Wei, Carley K M. Exploring patterns of identity usage in tweets: A new problem, solution and case study [C] //Proc of the 25th Int Conf on World Wide Web. New York: ACM, 2016: 401-412

[44] Dong Yuxiao, Yang Yang, Tang Jie, et al. Inferring user demographics and social strategies in mobile social networks [C] //Proc of the 20th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2014: 15-24

[45] Li Jiwei, Ritter A, Hovy E. Weakly supervised user profile extraction from Twitter [C] //Proc of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2014: 165-174

[46] Ikeda K, Hattori G, Ono C, et al. Twitter user profiling based on text and community mining for market analysis [J]. Knowledge-Based Systems, 2013, 51(1): 35-47

[47] Tang Cong, Ross K, Saxena N, et al. What's in a name: A study of names, gender inference, and gender behavior in Facebook [C] //Proc of the Int Conf on Database Systems for Advanced Applications. Berlin: Springer, 2011: 344-356

[48] Pazzani M, Billsus D. Learning and revising user profiles: The identification of interesting Web sites [J]. Machine Learning, 1997, 27(3): 313-331

[49] Chan P K. Constructing Web user profiles: A non-invasive learning approach [C] //Proc of the KDD Workshop on Web Usage Analysis and User Profiling. New York: ACM, 1999: 39-55

[50] Soltysiak S J, Crabtree I B. Automatic learning of user profiles—Towards the personalization of agent services [J]. BT Technology Journal, 1998, 16(3): 110-117

[51] Makazhanov A, Rafiei D, Waqar M. Predicting political preference of Twitter users [J]. Social Network Analysis and Mining, 2014, 4(1): 193:1-193:15

[52] Figueiredo F, Ribeiro B, Almeida J M, et al. TribeFlow: Mining and predicting user trajectories [C] //Proc of the 25th Int Conf on World Wide Web. New York: ACM, 2016: 695-706

[53] Wu Le, Ge Yong, Liu Qi, et al. Modeling users' preferences and social links in social networking services: A joint-evolving perspective [C] //Proc of the 30th AAAI Conf on Artificial Intelligence. Menlo Park: AAAI, 2016: 279-286

[54] Wu Fang, Huberman B A. Novelty and collective attention [J]. Proceedings of the National Academy of Sciences, 2007, 104(45): 17599-17601

[55] Dezsö Z, Almaas E, Lukacs A, et al. Dynamics of information access on the Web [J]. Physical Review E, 2006, 73(6): 066132

[56] Yang J, Leskovec J. Modeling information diffusion in implicit networks [C] //Proc of the 10th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2011: 599-608

[57] Yan Rui, Tang Jie, Liu Xiaobing, et al. Citation count prediction: Learning to estimate future citations for literature [C] //Proc of CIKM 2011. New York: ACM, 2011: 1247-1252

[58] Yan Rui, Huang Congrui, Tang Jie, et al. To better stand on the shoulder of giants [C] //Proc of JCDL 2013. New York: ACM, 2013: 51-60

[59] Livne A, Adar E, Teevan J, et al. Predicting citation counts using text and graph minin [C] //Proc of the Conf 2013 Workshop on Computational Scientometrics: Theory and Applications. New York: ACM, 2013

[60] Shen Huawei, Wang Dashun, Song Chaoming, et al. Modeling and predicting popularity dynamics via reinforced poisson processes [C] //Proc of 28th AAAI Conf on Artificial Intelligence. Menlo Park: AAAI, 2014: 291-297

[61] Pobiedina N, Ichise R. Citation count prediction as a link prediction problem [J]. Applied Intelligence, 2016, 44(2): 252-268

[62] Tang Jie, Zhang Jing, Yao Limin, et al. ArnetMiner: Extraction and mining of academic social networks [C] // Proc of the 14th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2008: 990-998

[63] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022

[64] Deerwester S, Dumais S T, Furnas J W, et al. Indexing by latent semantic analysis [J]. Journals of the American Society for Information Science, 1990, 41(6): 391-407

[65] Hofmann T. Probabilistic latent semantic indexing [C] // Proc of ACM SIGIR 1999. New York: ACM, 1999: 211-218

[66] Griffiths T, Steyvers M. Finding scientific topics [J]. Proceedings of National Academy of Sciences. 2004, 101(1): 5228-5235

[67] Minka T, Lafferty J. Expectation propagation for the generative aspect model [C] //Proc of UAI 2002. San Francisco: Morgan Kaufmann, 2002: 352-359

[68] Wei Xing, Croft W B. LDA-based document models for ad-hoc information retrieval [C] //Proc of the 29th SIGIR 2006. New York: ACM, 2006

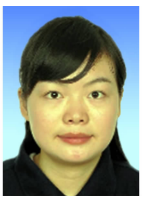
[69] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks [C] //Proc of the 30th Int Conf on Machine Learning. New York: ACM, 2013: 1310-1318

[70] Xiao Shuai, Yan Junchi, Yang Xiaokang, et al. Modeling the intensity function of point process via recurrent neural networks [C] //Proc of the 31st AAAI Conf on Artificial Intelligence. Menlo Park: AAAI, 2017: 1597-1603

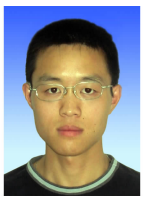
[71] Zhuang Yan, Li Guoliang, Fen Jianhua. A survey on entity alignment of knowledge base [J]. Journal of Computer Research and Development, 2016, 53(1):165-192 (in Chinese)
(庄严，李国良，冯建华. 知识库实体对齐技术综述[J]. 计算机研究与发展，2016，53(1)：165-192)

[72] Liu Zhiyuan, Sun Maosong, Lin Yankai, et al. Knowledge representation learning: A review [J]. Journal of Computer Research and Development, 2016, 53(2): 247-261 (in Chinese)
(刘知远，孙茂松，林衍凯，等. 知识表示学习研究进展[J]. 计算机研究与发展，2016，53(2)：247-261)

[73] Liu Jiao, Li Yang, Duan Hong, et al. Knowledge graph construction techniques [J]. Journal of Computer Research and Development, 2016, 53(3): 582-600 (in Chinese)
(刘峤，李杨，段宏，等. 知识图谱构建技术综述[J]. 计算机研究与发展，2016，53(3)：582-600)



Yuan Sha, born in 1989. PhD. Her main research interests include scholar profiling, machine learning and social network.



Tang Jie, born in 1977. PhD, associate professor and PhD supervisor in Tsinghua University. IEEE Senior Member, ACM Senior Member, CCF Distinguished Member. His main research interests include network theories, data mining methodologies, machine learning algorithms and semantic Web technologies.

Gu Xiaotao, born in 1994. PhD candidate in University of Illinois at Urbana-Champaign. His main research interests include information extraction, data mining and natural language understanding.