doi:10.3772/j.issn.2095-915x.2019.03.006

开放科学
（资源服务）
标识码
（OSID）

面向学术文献的学者兴趣标签识别方法

谢鹏

首都图书馆　北京　100021

摘要：学术文献是科学进步与发展的载体，各种元数据信息包括作者、论文、期刊以及这些实体之间的关系，具有重要的价值，如何精准构建学者用户画像是一个具有挑战性的问题。早期的用户画像相对简单，区分度以及可用性都不高。本文在“2017 开放学术精准画像大赛”TASK3 的真实数据上，提取学者与期刊的关系和学者与论文的关系，设计关系模型并采用 LSI 降维技术以及文本相似度计算，对学者兴趣标签进行识别与评估，并进行数据可视化分析。实验结果表明，使用本文提出的方法可准确有效的识别学者兴趣标签，准确率为 P@1=92%、P@2=94%、P@3=98%。

关键词：用户画像；兴趣标签识别；LSI
中图分类号：G35

Recognition of Scholar Interest Tag for Academic Literatures

XIE Peng

Capital Library of China, Beijing 100021, China

Abstract: literature is recognized as the carrier of scientific progress and development. Various metadata information, including the author, thesis, press, and even the relationship between these entities, is of great value. How to construct user profile for academic users exactly is a challenging issue. The early user profile is relatively simple, with little distinction and usability. Based on the real data set of task 3 in “2017 Open Academic Data Challenge”, we extract the relationship between scholars and press, and the relationship between scholars and thesis to design the relationship model. And then, use the LSI dimensionality reduction technology and the similarity calculation of text to recognize the scholar’s interest. The recognized interests are evaluated and do data visualization analysis. The experimental results show that the method proposed based on the information of the press and thesis in this paper can effectively and accurately recognize scholars interest labels. And the accuracy is P@1=92%, P@2=94%, P@3=98%.

Keywords: User profile; interest label recognition; LSI

作者简介：谢鹏（1980-），学士，馆员，研究方向：图书情报、数字图书馆，E-mail：xp@clcn.net.cn。

doi:10.3772/j.issn.2095-915x.2019.03.006

引言

学术文献记录着科学的进步与发展，其价值得到了各个领域的重视和认可。其中的学术信息，包括作者、论文、期刊以及这些实体之间的关系，对研究领域以及企业都具有独特的价值。所以如何精准的从海量学者信息中准确识别学术信息并精准构建学者用户画像变成一个具有挑战性的问题。精准的学术信息提取以及识别并构建学者用户画像不仅可以丰富论文的共享方式，还能有效帮助研究人员进行学术检索与交流，缩短科研成果产业化周期。

用户画像又称人群画像，是根据用户人口统计学信息、社交关系、偏好习惯和消费行为等信息而抽象出来的标签化画像，其核心工作是给用户贴“标签”。早期的用户画像相对简单，类似于个人档案信息，区分度以及可用性都不高。随着步入大数据时代，一些用户画像技术涌现出。例如用户在搜索引擎的搜索栏中描述其信息需求，会显示一系列的搜索结果，用户点击其感兴趣的一个或多个链接，当其信息需求未得到满足时，会用不同的方式重新表示先前的查询[1]。通过对人群基本属性、行为习惯、行业价值等多个维度的信息进行综合分析，可以精准的进行目标受众的画像和定位，实现基于大数据的精准营销。

本文提出基于期刊信息以及文献内容信息识别学者兴趣标签并对采集的数据进行可视化分析。主要工作如下：

1) 学者兴趣标签的识别。根据学者的学术相关信息，提取学者与期刊的关系和学者与论文的关系，并设计关系模型。采用 LSI 降维技术以及文本相似度计算，计算得到学者各个兴趣的概率。调节权重将两个模型结果融合，进而得到最终概率分布，并按照概率大小排序识别的兴趣标签。

2) 识别结果的评估。选取 Top3 的识别标签，与学者个人主页中的 Introduction 比较，使用交叉验证的方法计算识别准确率。

3) 数据可视化分析。统计所获取的学者兴趣标签数据信息，使用 Echarts 开源数据可视化工具进行数据可视化分析，设计学者个人画像并直观体现数据之间的关系。

1 国内外研究现状

随着互联网上大量的信息和服务可用性的提升，个性化的设计已成为帮助用户搜索、过滤和选择感兴趣项目的重要内容。个性化被描述为能够根据用户的偏好和行为知识为个人提供量身定制的内容和服务[2]。用户画像对于识别用户的兴趣、行为或其他都可以用于形成个性化的特征。用户画像的构建，大体上分为用户画像方向或者分类体系的建立、用户数据收集、用户标签或指数建模研究三个步骤。对此目前已经提出了一些可行的技术方法。

有很多应用的研发方向是为寻找用户的兴趣以及专业知识，而其最重要的任务之一是个性化搜索和推荐。目前有大量的研究[3-4]一直致力于挖掘用户的潜在兴趣或专业知识，利用用户的个人偏好在不同领域构建推荐系统，如网络搜索，网页内容，评分系统[5]以及社交媒体[4,6]。

可以通过直接用户参与或隐式的通过用户活动自动监控来显式的收集信息[7]。用户画像

可以由用户自己来明确定义，例如在用户注册某种服务期间。网络用户行为分析，即从海量网络数据中挖掘其潜在的、有价值的数据[8]。现已进行了关于获得用户画像和用户画像模型的类型[9]（例如，基于内容的、分层的、语义的用户画像模型）的各种最先进技术的调研。刘鹏提出，网络用户行为分析是指对用户上网数据进行统计分析并从中发现网络用户行为的规律，而这种规律可以用某些特征公式表示，也可以通过可视化用图表体现出来其特征[10]。郝增勇在其基于 Hadoop 用户行为分析系统设计与实现一文中，总结了当下用户行为分析的主要研究方法有：用户特征分析，关联分析，分类与预测，异常分析以及 TopN 分析[11]。

用户画像可以表示为关键词的向量。有一个常用的方法用于表示基于用户消费或生产内容的用户画像，即作为术语（关键字）的加权矢量，也被称为词袋（BOW）模型[12]。在这个模型中，用户画像的术语通常是根据信息检索 tf×idf 方法[13]进行加权的。大多数先前的技术中，关键词的权重使用 tf×idf 方案或其变体进行分配[14]。其他方法从开放式目录管理系统（ODP）分层结构中提取分类标准以表示用户画像[15]。

在社交媒体系统中，人口统计信息通常包括姓名、年龄、性别和位置，为创建信息个性化的用户模型提供了重要的数据资源，例如用户画像中对于用户流失行为的研究。郭明结合移动流失客户的实际数据分布特征，进行过度抽样建立了基于决策树 C4.5 的客户流失预测模型[16]；而 Bayes 神经网络算法也得到广泛使用，另外还有一些其他算法也应用于客户流失预测模型的建立与研究，如 K 近邻序列算法、生存分析法等，以及基于结构风险最小化理论的支持向量机 (SVM) 模型也大量应用于客户流失预测研究中。如武帅、王雄和段云峰提出了基于 SVM 的电信客户流失预测模型[17]，实验结果表明，该模型具有较好的预测性和稳定性。

另有一些用户画像的研究从用户行为的各个角度出发，如用户需求、用户参与，用户价值，用户心理等各个角度进行了展开。随后，学者们以手机为研究对象，以更精确的用户空间行为数据进行实证探究[18]。研究表明，用户行为在时间和空间上存在着重复性和可预测性。为构建高质量的用户主题画像以改善用户体验并为诸多网络个性化搜索、推荐系统、专家挖掘和社区检测等重要应用提供动力，有两种主要的方法。一种方法是试图寻找可能描述用户的潜在因素。例如，在社交媒体的用户帖子上运行 Latent Dirichlet Allocation（LDA）模型可以揭示用户感兴趣的主题[19]；另一个方法线索旨在鼓励社交媒体用户直接评估彼此的兴趣和专业知识，从而为用户主题专业人士提供一个部分的视角。例如，LinkedIn 用户可以为自己的个人资料选择技能标签，并可以在其他人的个人资料上认可这些标签[20-21]。总而言之，这种人群贡献的标签知识可以被看作是捕获用户兴趣和专业知识的可靠依据。

2 学者兴趣标签识别模型

学者兴趣标签识别包括基于期刊信息的学者兴趣标签识别与基于文献内容信息的学者兴

doi:10.3772/j.issn.2095-915x.2019.03.006

趣标签识别两个模型，将两个模型结果融合得到最终识别结果，并对该识别结果进行数据可视化分析以及识别准确率评估。学者兴趣标签识别的整体设计框架如图 1 所示。

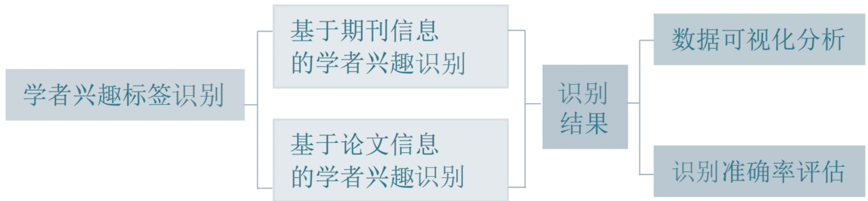

图 **1** 学者兴趣标签识别整体设计框架图

2.1 元数据抽取

论文数据集用“2017 开放学术精准画像大赛”TASK3 所提供的论文数据集 paper.txt，本数据集由 Aminer 提供，共包含 1,609,106 名学者的论文及标签数据，包含了学者截至 2016 年所有论文的数据。学者兴趣数据集用“2017 开放学术精准画像大赛”TASK2 所提供的论文数据集 training.txt，本数据集由 Aminer 提供，共包含 15000 位学者的兴趣标签数据（收集时间：2017 年 7 月 10 日），包含 6000 个学者。学者姓名数据集包含 6000 个学者姓名，即待识别兴趣标签的学者。

从数据集中提取出“作者 - 合作者”、“作者 - 序号 - 引用”、“作者 - 兴趣”、“作者 - 期刊”、“序号 - 作者 - 文章”、“作者 - 文章”的关系。字典中出现的符号如下：Author 代表作者，Cooperator 代表合作者，Press 代表期刊，Paper 代表论文，Index 代表序号，Cite 代表所引用的论文。包括“Author-Press”关系、“Author-Cooperator”关系、“Author-Index-Cite”关系、“Index-Author-Paper”关系、“Author-Paper”关系以及“Author-Interest”关系。

2.2 学者 – 论文 – 期刊关系网络构建

2.1 中抽取的学者与论文的关系以及学者与期刊的关系可表示为图 2 所示形式，Author1 为待识别学者，Author2 与 Author3 为学者兴趣数据集中已知兴趣的学者，Author 1 与 Author 2 同时发表过 Paper1、Paper2 与 Paper3，Author 1 与 Author 3 同时发表过 Paper4，Paper1 与 Paper2 发表于 Press1，Paper3 发表于 Press2，Paper4 发表于 Press3。

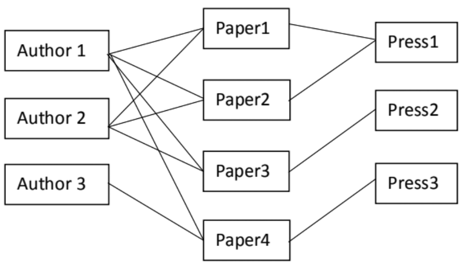

图 **2** 学者 - 论文 - 期刊关系网络图

2.3 基于期刊信息的学者兴趣标签识别

首先生成“Press-Interest”关系，期刊通常收录某一领域的论文，具有一定的兴趣集合。学者将论文投到与自己研究兴趣相近的期刊。模型先计算期刊的兴趣分布，然后根据待识别学者投稿期刊的信息，识别学者的兴趣。

其次，进行学者对期刊偏好的计算。利用

公式 1 计算学者 i 的每个 Press 出现的频率。

$$w_i = [\frac{C(P_j, P_{A_i})}{|P_{A_i}|}]^3 \qquad （公式 1）$$

其中，P_{A_k}代表学者 A_i 所发表文章的全部期刊集合，P_j 代表期刊 j，其中 $P_j \subseteq P_{A_k}$, C(m,n) 代表 m 在 n 中出现的次数。

最后，利用公式 2 计算基于期刊信息的学者兴趣。

$$I_{A_i-press} = \sum_{k=0}^{P} w_k * identify(I_m, S_{P_{A_k}}^I) \quad （公式 2）$$

其中，$I_{A_i-press}$代表基于期刊信息兴趣 I_m 可能为学者 A_i 兴趣的概率，I_m 代表兴趣 m，$S_{P_{A_k}}^I$代表“Press-Interest”集合，k 为期刊编号，P 为全部期刊中最末期刊编号，k ∈ [0,P]。

2.4 基于文献内容信息的学者兴趣标签识别

首先，生成“Interest-Paper”关系。有合作关系的作者一起发表同一篇论文，从其合作的论文中反映出学者与其合作者可能有着相同的研究兴趣。故，将学者发表的论文集合按照学者的兴趣进行分类。

其次，计算“Author-Paper”关系与“Interest-Paper”关系集合重合度。Auhtor_x 发表的论文集合与“Interest-Paper”关系中各个 Interest 包含的论文集合会有部分重复内容，重复内容越多代表该 Interest 为 Auhtor_x 的兴趣的可能性越大。

最后，将文本转换成 TF-IDF 来表示，利用 LSI 模型降为并利用工具 Gensim 对“Author-Paper”与“Interest-Paper”进行文本相似度计算，如公式 3 所示，计算基于文献内容信息的学者兴趣。

$$I_{Ai-paper} = \mathrm{Similarity}(S_{A_i}^T, S_{I_m}^T) \qquad （公式 3）$$

2.5 两种识别结果的融合

将基于期刊信息的学者兴趣标签识别结果以及基于文献内容信息的学者兴趣标签识别结果按照一定权重 a 和 b 进行融合。具体算法如图表 1 所示。

表 1 两种模型识别结果融合算法

算法1：两种模型识别结果融合算法
输入： 基于期刊的学者兴趣标签识别结果，基于论文的学者兴趣标签识别结果 输出： Top n 的interest名称，即为该学者的兴趣标签识别结果 Begin 利用计算公式：$I_m = a*I_{Ai-press} + b*I_{Ai-paper}$， 计算每位学者的每一个interest的概率 调整权重：与学者个人主页introduction中的interest进行对比，调整权重使结果更接近答案 排序：将概率从大到小进行排序，即概率越大的interest越有可能是该学者的兴趣。 End

3 实验及分析

3.1 数据集

论文数据集共包含 1,609,106 名学者截至 2016 年的论文及标签数据，数据格式如表 4-1 所示。每篇论文包含六个字段（有些论文某些字段缺失）。学者兴趣标签共 6000 个学者，不同学者之间用空行分隔，具体格式如表 2 所示。学者姓名数据集包含 6000 个学者姓名，即待识别兴趣标签的学者。

表 2 论文数据集格式

字段	含义
#index	编号
#*	标题
#@	学者，多名学者以逗号分开
#t	发表年份
#c	发表的会议、期刊
#%	包含多行，每一行代表一篇所引用论文的id

doi:10.3772/j.issn.2095-915x.2019.03.006

表 3 学者兴趣数据集格式

字段	含义
#name	学者的名字
#interests	学者的兴趣标签(每位学者有3个兴趣标签，逗号分隔)

3.2 学者兴趣标签识别

"Author-Interest"字典包含 6000 名学者，以及每位学者可能的全部兴趣的概率。基于期刊信息的学者兴趣标签识别结果样例如表 4 所示，基于文献内容信息的学者兴趣标签识别结果样例如表 4 所示。识别结果存为 csv 格式，方便后续数据可视化统计分析，最终学者兴趣标签识别结果格式样例如表 5 所示。

表 4 基于期刊信息的学者兴趣标签识别结果样例表

Name	Label	Probability
M. N. Doja	linear algebra	0.00020783530362074423
	data security	2.8380871292748687e-06
	information fusion	0.00027053051957266517
	computer algebra	7.379026536114659e-05
	graphical model	0.0004228749822619554
	optical flow	3.1859816806053364e-05

表 5 基于文献内容信息的学者兴趣标签识别结果样例表

Name	Interest	Probability
Sudeepta Mishra	computational physics	0.03430805206298828
	formal specification	0.07013462409377098
	integer programming	0.10374868139624596
	evaluation	0.01808186424896121
	artificial intelligence	0.1565066769719124
	social informatics	0.022443460812792183

表 6 学者兴趣标签识别结果格式样例

Author	Interest1	Interest2	Interest3
Trey Ideker	bioinformatics	computational biology	gene expression
Chongcheng Chen	remote sensing	data mining	geographic information system
Uday V. Shanbhag	game theory	optimization	decision analysis
Yi-Hsuan Yang	computer vision	feature extraction	audio signal processing

3.3 识别结果评估

共 6000 名学者，将其分为五组，每组 1200 人，从每组中任意 10 个人，找到其个人主页进行交叉验证。对五组评估结果求均值得到最终评估结果，如表 7 所示，评估结果为 P@1=0.92，P@2=0.94，P@3=0.98。

表 7 最终评估结果

Author	P@1	P@2	P@3
Group1	90%	90%	100%
Group2	90%	90%	100%
Group3	90%	100%	100%
Group4	90%	90%	90%
Group5	100%	100%	100%
Average	0.92	0.94	0.98

3.4 数据可视化分析

每位学者都有其兴趣列表，为更加直观体现每位学者的研究兴趣，采用 Echarts 数据可视化工具，为每位学者设计其研究兴趣的个人画像，标签越大则说明学者更注重于该领域的研究。如图 3 所示，学者 Zoran Kalafatic、Aaron Clauset、Gitta Kutyniok、Hessam Mahdavifar、Michal Sorel、Rodney M. Goodman 六位学者的 Top19 interest 展示。

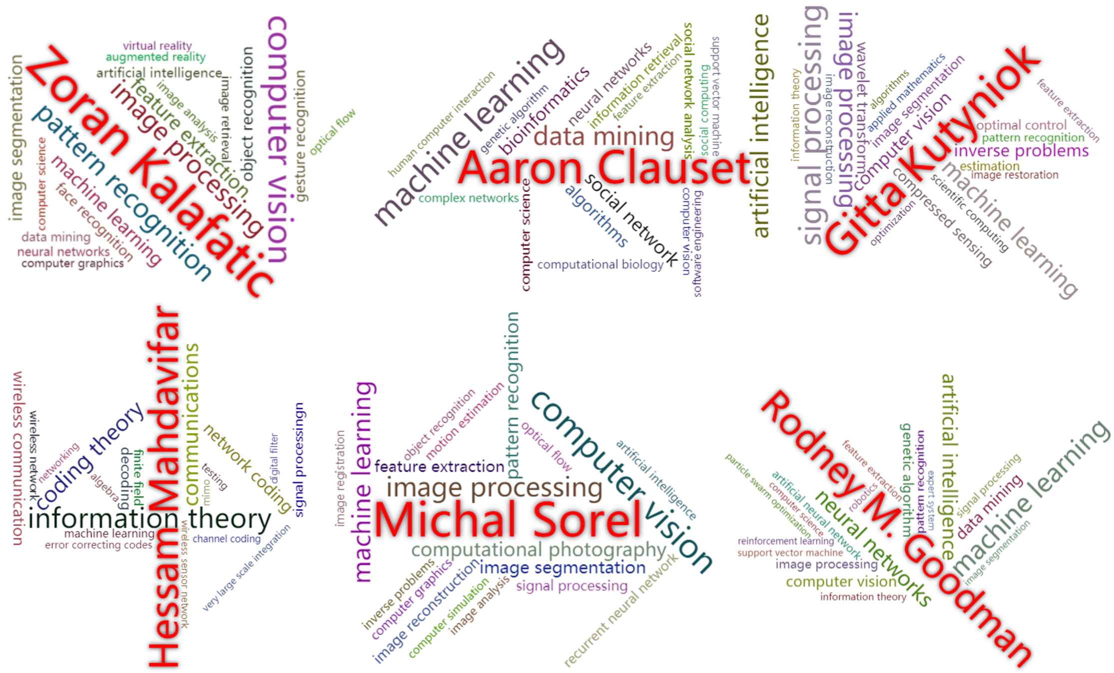

图 3 学者兴趣标签个人画像

4 结论

精准的学者用户画像构建可以高效的利用学术信息，不仅可以丰富论文的共享方式，还能有效的帮助研究人员进行学术检索与交流，以缩短科研成果产业化周期。为实现精准的从海量学者信息中有效的提取与识别学术信息并精准构建学者用户画像。本用户画像发现与挖掘技术借助“2017 开放学术精准画像大赛”提供的数据集，实现了对学者兴趣标签的精准识别。得到结论如下：

1) 本文根据数据集中提供的学者的学术信息，通过元数据抽取形成了学者 - 论文 - 期刊关系网络，并利用 LSI 降维技术以及文本相似度计算设计基于期刊信息识别学者兴趣标签和基于文献内容信息识别学者兴趣标签两个模型分别进行识别并对识别结果进行融合得到最终结果。最后采用交叉验证的方法对识别结果进行分组评估得到 P@1=0.92、P@2=0.94、P@3=0.98 的结果。实验结果表明，该识别方法可精准识别学者兴趣标签。

2) 识别信息的目的是为了更好的利用，本

doi:10.3772/j.issn.2095-915x.2019.03.006

文采用 Echarts 免费开源数据可视化工具从不同角度对获取的数据统计后进行数据可视化分析，更加直观的反映学者兴趣标签的分布。

虽然本文中提到的方法在解决以上提到的问题方面表现的很好，但还不是最优方案。可以借助机器学习和深度学习方法，现有的用户画像技术各有千秋，炫耀分别尝试后将识别结果进行对比，取其精华去其糟粕，然后在算法上优化并创新。研究学者用户画像的目的是为了提高学术信息的可用性，所以可以利用本文提到的方法将其进一步设计成一个学术搜索引擎，输入学者的姓名以及其所在机构，返回该学者的基本信息以及其兴趣标签，并且兴趣标签关联着与之有关的其他作者，这样形成一张学术信息网，使得学术信息更好的被发现于挖掘。

参考文献

[1] Jansen B J, Spink A, Saracevic T. Real life, real users, and real needs: A study and analysis of user queries on the web[J]. Information Process Management, 2000, 36(2):207-227.

[2] Gao M, Liu K, Wu Z. Personalisation in web computing and informatics: Theories, techniques, applications, and future research[J]. Information Systems Frontiers, 2010, 12:607–629.

[3] Guy I, Zwerdling N, Ronen I, et al. Uziel. Social media recommendation based on people and tags[C]. In Special Interest Grouop on Information Retrieval (SIGIR), 2010.

[4] Jiang M, Cui P, Liu R, et al. Social contextual recommendation[C]. In Conference on Information and Knowledge Management (CIKM), 2012.

[5] Jamali L, Lakshmanan. Recommendation in heterogeneous information networks using context dependent factor models[C]. In The International Conference of World Wide Web (WWW), 2013.

[6] Hong, L, Doumith A S, Davison B D. Co-factorization machines: modeling user interests and predicting individual decisions in twitter[C]. In International Conference on Web Search and Data Mining (WSDM), 2013.

[7] Ghorab M, Zhou D, O'Connor A. Personalised information retrieval: survey and classification[J]. User Modeling and User-Adapted Interaction, 2013, 23(4):381–443.

[8] 任思颖 . 基于大数据的网络用户行为分析 [D]. 北京 : 北京邮电大学 , 2014.

[9] Gauch S, Speretta M, Chandramouli A, Micarelli A. User profiles for personalized information access[C]. In The Adaptive Web, volume 4321 of Lecture Notes in Computer Science. Berlin, Heidelberg, 2007.

[10] 刘鹏 . 网络用户行为分析的若干问题研究 [D]. 北京 : 北京邮电大学 , 2010.

[11] 郝增勇 . 基于 hadoop 用户行为分析系统设计与实现 [D]. 北京 : 北京交通大学 , 2014.

[12] Chen L, Sycara K. Webmate: a personal agent for browsing and searching. In ACM AGENTS '98, New York, NY, USA. 1998.

[13] Manning C D, Raghavan P, Schutze H. Introduction to Information Retrieval[M]. Cambridge University Press, 2008.

[14] Vallet D, Cantador I, Jose J M. Personalizing web search with folksonomy-based user and document profiles[C]. In Proceedings of ECIR'10. Springer-Verlag, 2010: 420–431.

[15] Speretta M, Gauch S. Personalized search based on user search histories[C]. In Proceedings of IEEE/WIC/ACM WI'05. IEEE Computer Society, 2005: 622–628.

[16] 郭明 . 基于决策树的客户流失分析 [J]. 广东通信技术 , 2004, 24(11):37-40.

[17] 武帅 , 王雄 , 段云峰 . SVM 在移动通信客户流失预测中的应用研究 [J]. 微计算机信息 , 2007, 23(4):163-165.

[18] 汪秉宏 , 韩筱璞 . 人类行为的动力学与统计力学研究 [J]. 物理 , 2010, 39(1):28-37.

[19] Majumder A, Shrivastava N. Know your personalization: learning topic level personalization in online services[C]. In The International Conference of World Wide Web (WWW), 2013.

[20] Bhattacharya P, Ghosh S, Kulshrestha J, et al.. Deep twitter diving: Exploring topical groups in microblogs at scale[C]. In the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW), 2014.

[21] Ghosh S, Sharma N, Benevenuto F, et al. Cognos: crowdsourcing search for topic experts in microblogs[C]. In Special Interest Grouop on Information Retrieval (SIGIR), 2012.