文章编号:1007-5321(2020)01-0122-07 **DOI**:10.13190/j.jbupt.2019-020

社交网络用户身份关联及其分析

孙 波, 张 伟, 司成祥
(国家计算机网络应急技术处理协调中心, 北京 100029)

摘要: 同一用户在不同社交平台注册账号,使得用户数据分散于多个平台,且这些数据不全面、不可靠、利用率低. 通过分析这些跨平台的数据,发现不同账户对应同一用户的真实身份,使跨平台用户身份关联,以构建详细的用户画像、推荐系统、跨社交网络的链接预测等. 从国内外身份关联技术的研究现状出发,介绍了用户身份关联及分析框架,整理了身份数据采集标准和社交网络数据集;分析了近几年用户身份关联技术,并归纳了身份关联评价指标,阐述了基于身份关联的社交网络数据挖掘及分析框架;最后对身份关联中的研究难点及热点进行了讨论和展望.

关 键 词: 跨平台; 身份关联; 身份识别
中图分类号: TP391.3 **文献标志码**: A

Social Network User Identity Association and Its Analysis

SUN Bo, ZHANG Wei, SI Cheng-xiang
(National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China)

Abstract: The same user registers accounts on different social platforms, which makes user data scattered across multiple platforms, and these data are incomplete, unreliable and low utilization. By using these cross-platform data to discover the real identity of the same user corresponding to different accounts, cross-platform user identity association plays an important role in building detailed user profiles, recommendation systems, cross-social network link prediction and other cross-platform applications. Starting from the research status of identity association technology at home and abroad, the framework of user identity association and analysis is introduced, and the standards of identity data acquisition and social network data sets are collated. Subsequently, the technology of user identity association in recent years is analyzed and the evaluation index of identity association is summarized, and the social network data mining and analysis based on identity association is expounded. Finally, the research difficulties and hotspots of identity association are discussed and prospected.

Key words: cross-platform; identity association; identification

随着通信网络的发展,移动应用受到越来越多的关注. 豆瓣、新浪微博、微信、天涯论坛、Facebook、Twitter 等国内外著名网络应用的影响力不断提高,并深入人们的日常生活中. 研究人员基于这些网络应用,开展影响力分析、用户行为模式研究、用户偏好获取、移动推荐等应用. 然而,由于公司利益和用户隐私保护的需求,多数应用研究仅限于单一的数据源,这使得当前的研究只能分析用户片面的行为模式及偏好,难以全面刻画用户画像. 如何高效地融合多个数据源的用户信息,提供更加全面的用户

收稿日期: 2019-01-30
基金项目: 国家重点研发计划项目(2018YFB0804800)
作者简介: 孙 波(1973—), 男, 正高级工程师.
通信作者: 司成祥(1982—), 男, 高级工程师, E-mail:sichengxiang@cert.org.cn.

信息，将有利于更全面、准确地开展研究，因此，身份关联技术受到越来越多的关注. 近几年在国际权威期刊和顶尖会议上，如 TKDE[1-2]（IEEE trans on knowledge and data engineering）（CCF A 刊）、SIGMOD[3]（ACM conference on management of data）（CCF A 会）、SIGKDD[4-5]（ACM knowledge discovery and data mining）（CCF A 会）、VLDB[6]（international conference on very large data bases）（CCF A 会）、CIKM[7-8]（ACM conference on information and knowledge management）（CCF B 会）、WWW[9]（the web conference）（CCF A 会）等，都有相关的报道. 在国内，对身份关联的研究主要集中在社交网络用户的识别[10]和复杂网络节点匹配[11]上. 例如，周小平等[12]对国内外相关文献进行了综述. 李晓菲等[13]提出了一种跨平台的微博社区账户匹配方法等. 但总体而言，对于网络身份关联技术的研究还处于探索阶段.

1　社交网络用户身份关联及分析框架

对当前用户身份关联领域的研究进行了归纳，社交网络用户身份关联及分析框架如图 1 所示. 框架分为 4 个部分，自底向上分别对应 2 ~ 5 节的内容.

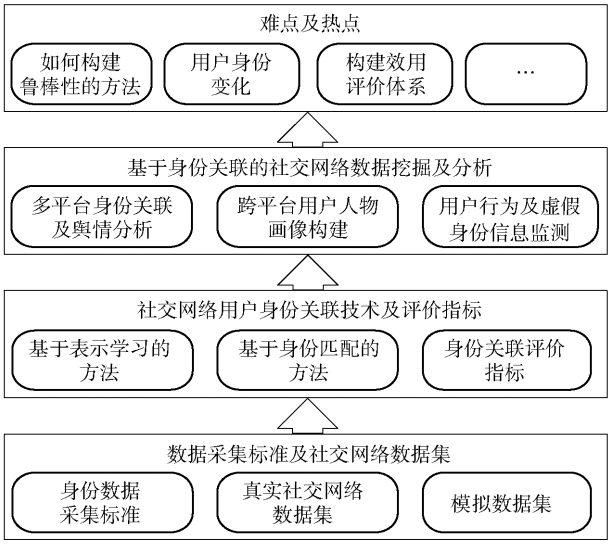

图 1　社交网络用户身份关联及分析框架

在数据采集标准及社交网络数据集部分中列举了一些身份数据采集标准，整理了一些真实社交网络数据集及模拟数据集.

社交网络用户身份关联技术及评价指标部分主要介绍基于表示学习的方法和基于身份匹配的方法；整理了部分身份关联评价指标，用于评价身份关联的实际效果.

基于身份关联的社交网络数据挖掘及分析部分包含多平台身份关联及舆情分析、跨平台用户人物画像构建和用户行为及虚假身份信息监测. 最后总结了用户身份关联中存在的问题和热点.

综上所述，该框架在逻辑上以数据采集标准及社交网络数据集为基础，深入分析社交网络用户身份关联技术，并通过身份关联评价指标衡量关联效果；对用户身份关联的应用情况进行了总结，主要体现在基于身份关联的社交网络数据挖掘及分析部分；最后，对社交网络用户身份关联及分析中存在的难点以及未来研究的热点进行讨论和展望.

2　数据采集标准及社交网络数据集

2.1　身份数据采集标准

在研究用户关联的过程中，获取的身份数据集通常具有多平台、异构性、大规模、多模态等特征，需要一定的标准进行约束. 目前我国在数据采集方面的相应标准包括 YD/T 2673—2013 面向舆情分析的互联网数据采集与交换格式定义、SJ/T 11615.1—2016 网络数据采集分析软件规范、YD/T 2405—2015 互联网数据中心和互联网接入服务信息安全管理系统接口规范、YDB 147—2014 互联网舆情监测与分析系统框架等.

2.2　真实社交网络数据集

目前，许多平台都可以采集到丰富的社交网络数据，包括微信、新浪微博[12]、豆瓣网①、人人网[12]、Facebook[14]、Twitter[15]等社交平台；大众点评[16]、京东商城[17]等生活及电子商务平台；163 邮箱、Gmail 邮箱等邮箱网络. 手机、个人电脑、平板电脑、浏览器等用户入网设备/工具都可实现数据收集.

研究者从真实网络中按照一定比例抽取出实验用户及相应的用户关系，抽样出的网络存在一定的重叠. 真实数据集合的优势在于存在一定数量的用户属性数据，也更接近真实情况.

2.3　模拟数据集

除上述真实数据集外，还有一些网络或模型可用于构造模拟数据，这些网络或模型主要包括 Erdös-Rényi 网络（ER，Erdös-Rényi）[18]、Watts-Strogatz 网络（WS，watts-strogatz）[19]、Barabási-Albert 网

① https://beijing.douban.com/

络(BA,barabási-albert)[20]、递推矩阵模型(R-MAT,recursive matrix)[21]等. 例如,Zhou 等[12]利用 ER、WS、BA 这3种网络模型规则分别生成人工网络,并抽取出一对模拟社交网络,在模拟网络上对已知的关联账号技术进行实验.

3 社交网络用户身份关联技术及评价指标

从模型角度开展分析,将身份关联技术划分为基于表示学习的方法和基于身份匹配的方法,其中基于表示学习的方法,是指利用用户数据将身份映射表示在统一空间,从而发现身份是否关联,该方法建立在对用户信息的深度提取和用户特征的准确表示上. 而基于身份匹配的方法,是指利用身份关联的一些特征信息,例如头像、姓名、注册地点、上网内容或社会关系等信息,计算用户间的相似度,估计关联身份的概率,从而实现身份关联.

3.1 基于表示学习的方法

用户数据因隐私保护通常呈碎片化、不一致或者不可靠的特点. 因此利用表示学习的方法准确表示用户特征成为有效方法,目前较新的研究内容主要归于这类方法,可分为以下4种.

1) 基于网络结构与嵌入式的方法

使用基于网络结构的用户识别涉及网络嵌入的内容,旨在学习网络中的潜在表示. 大多数需要给定需要识别的用户的先验知识,通过嵌入方式[22]或构建嵌入模型[23]实现.

2) 无监督式的方法

在难以获得用户的先验知识的情况下,通常使用无监督式的方法. 例如,借助好友关系无监督地进行建模[24-25],获取特征向量,得到匹配用户.

3) 结合网络结构与其他信息的方法

单独基于网络结构具有一定的局限性,可将网络结构与其他信息结合起来. 例如结合具有网络结构的用户画像属性与社会链接[26],结合时间维度上的用户行为轨迹和用户的核心社交网络结构[3]等.

4) 基于特征建模的方法

为了更加了解用户的偏好特征,一种方法侧重于对用户信息或特征建模. Deng 等[27]通过注册信息、评论等计算用户相似度并线性加权,对用户兴趣建模. Mu 等[28]提出批处理模型和在线模型,模拟用户及其投影到不同的社交平台之间的关系.

综上所述,基于表示学习的方法从身份数据中挖掘深层次的特征,精准刻画身份实现关联. 表1对上述几种方法进行对比分析. 各个方法在一定程度上可以弥补各自的不足. 因此尝试将不同方法之间的组合是一个不错的选择.

表1 基于表示学习的方法对比分析

方法	描述	优缺点
基于网络结构与嵌入式	学习网络表示,嵌入特征增强识别	缓解稀疏性,增强表示,但依赖先验知识
无监督式	通常在无先验知识的情况下使用	减少先验知识影响,但普适性差,效果不稳定
结合网络结构与其他信息	将网络结构与其他方法结合提高识别效果	普适性强,缓解信息不足的影响,但无统一的研究框架
基于特征建模	侧重于对用户信息或特征建模	精确获取用户偏好,但易产生稀疏问题

3.2 基于身份匹配的方法

相对传统的研究中大量的方法都归于这类,利用身份相关的信息中相对独特和稳定的数据,实现相似性的计算,有效识别比较明显的关联身份,可分为以下6种.

1) 基于用户名与用户头像的方法

一类方法[4,7]认为账号的起名有一定的独特性,基于用户名发现关联账号. Zafarani 等[4]利用支持向量机(SVM, support vector machine)等监督学习的方法学习用户起名特征,从而识别关联用户. 也存在通过人脸识别算法计算用户头像相似性来发现关联账号的研究[29].

2) 基于多种属性的方法

为了提高关联的准确性,可借助多个用户属性,通过计算相似度确定用户身份是否匹配[30-31]. 例如 Iofciu 等[30]利用用户的多个属性刻画用户标签,计算用户向量之间的相似度发现关联账号. Motoyama 等[31]利用词袋模型从多种属性文本中提取出用户特征,进而计算用户相似度.

3) 基于用户创造内容的方法

用户创造的内容通常反映其意识变化和行为模式,且难以被伪造,可用于身份关联. 例如,通过分析用户书写内容风格识别关联用户[32-33]. 也可以通过考虑将用户创造内容的空间位置、发布时间以及用户成文风格作为识别依据[8-9].

4) 基于拓扑结构的方法

若假设2种平台的用户有较高的重叠度,这时将账号关联问题转化成为一种重叠身份识别问题,

主要利用账号的拓扑结构来分析. 在这类方法中，需要知道一些关联账号的信息作为种子[11,34]，以此衡量未知关联身份之间的相似性.

5）基于关联节点与相似度的方法

基于身份匹配对多个平台间的用户进行身份关联还会涉及关联节点的获取和对相似度的计算[35]，通过相似度与设定阈值的比较确定关联用户. 另外，也涉及节点到其他节点的度和距离信息[36].

6）结合社会关系与其他信息的方法

Zhou 等[1]根据研究的社交平台发现，67.5% 的新浪微博用户拥有人人网账户. Zhou 等[1]提出一种基于好友关系的社交网络用户挖掘方法，为所有候选用户匹配对计算匹配度. Yu[37]将关联账号识别任务转换为图的权重匹配问题，通过内容和社会关系计算节点间的相似度，得到关联的用户.

综上所述，基于身份匹配的每种方法中利用不同的元素或者属性进行用户身份关联. 但是也正是由于基于身份匹配的方法对数据的依赖性，使得数据的数量和质量的变化对实际效果影响较大，如何通过其他辅助数据缓解数据稀疏性的问题也是研究的重点之一. 从表 2 中对各方法的对比分析中也可以看出这一点.

表 2　基于身份匹配的方法对比分析

方法	描述	优缺点
基于用户名与用户头像	通过起名特性和头像匹配用户	利用起名和头像特点，但数据少且依赖数据
基于多种属性	利用多种用户属性，如标签等	缓解数据不足问题，但易受数据异构性影响
基于用户创造内容	利用创造内容反映的行为模式、特征	发现用户行为特征，缓解属性造假的影响，但因隐私保护，信息稀疏
基于拓扑结构	适用于平台之间的用户重叠识别	利用拓扑结构，但扩展能力差
基于关联节点与相似度	主要涉及关联节点的获取和相似度的计算	发现节点特征及节点之间的关系，但普适性差，对应用领域要求高
结合社会关系与其他信息	将稳定的社会关系与其他属性结合，提高精准度	缓解关系先验知识不足带来的影响，但效果提升不明显

3.3　身份关联评价指标

账号关联技术主要指标分为账号关联精度和响应时间. 目前，账号关联精度的主要测量指标包括准确率、召回率和 F1-Measure 等[12]. 其中准确率（P）是指所有被正确发现的关联账号数量占实际被发现的关联账号数量的比例，如

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{1}$$

其中：N_{TP}指被发现的正确账号关联用户个数，N_{FP}是指被错误判定为关联账号的非关联账号个数. 召回率（R）[12]指所有被正确发掘的关联账号占所有真实的关联账号的比例，如

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{2}$$

其中 N_{FN}是指被误判为非关联用户的关联用户. F1-measures 的定义为

$$F_1 = \frac{2PR}{P + R} = \frac{2N_{TP}}{2N_{TP} + N_{FP} + N_{FN}} \tag{3}$$

响应时间是指在相同的实验数据下，算法运行的时间. 通常数据规模越大，响应时间越长，而在相同的数据规模下，算法效率越高，响应时间越短.

4　基于身份关联的社交网络数据挖掘及分析

4.1　多平台身份关联及舆情分析

多平台用户身份数据中包含大量的用户意见或反馈信息，可以用于网络舆情分析. 对于舆情分析来讲，得益于用户身份关联，得到相对比较全面的用户信息. 随后就可以根据用户的各个账号信息，分析用户的意识型态，发现相关的意见倾向，客观反映各个平台的用户的舆情状态. 一个强大的舆情分析系统包括采集、处理、分析和加工过程[38]，通过多平台用户身份关联，将用户各方面数据搜集整理，得到较为准确的数据，从而进行后续分析，如跨媒体多源的舆情分析系统架构[39]等.

4.2　跨平台用户人物画像构建

通过账号关联发现用户在多个平台上对应的多个身份，从而根据用户在每个平台上的有效信息，刻画用户人物画像，借助社会化网络信息能够进一步提高用户满意度[40]. 此外，深度学习技术的广泛应用使得人物画像的构建逐渐摆脱单一平台数据的束缚. 例如通过结合深度学习来提高标注的性能[41]，引入卷积神经网络（CNN，convolutional neural networks）的方法. Chiu 等[42]结合使用循环神经网络（RNN，recurrent neural network）和 CNN，在特征获取方面有更加明显的进步. 另外，也存在基于语义发现人物偏好标签的研究[43].

4.3　用户行为及虚假身份信息监测

移动网络的发展丰富了用户的日常生活，但是

网络上出现大量虚假信息,影响了人们的决策甚至日常生活. 因此,亟需进行虚假信息监测. 根据使用数据的角度,可分为以下2种方法.

1) 基于文本分析的方法包括语法分析和语义分析. 语法分析主要通过提取词袋、词性等特征来实现[44-45]. 语义分析的方法是抽取语义层面的文本特征并加以分析的方法[45].

2) 基于数据风格的分析方法主要利用用户用词特征或句法特征来实现[45-46]. 基于应用风格的分析方法主要根据用户在不同平台上的不同行为表现,或者不同应用平台会有不同的风格和信息内容,判断用户行为信息的可信度[47-48].

5 难点及热点

目前,身份关联方面遇到的技术难点有:如何构建鲁棒性的方法;如何在账号关联的同时保护用户隐私安全;如何设计无关联用户知识的方法;如何应对真实环境中用户身份变化等.

在未来发展中,身份关联的研究热点主要包括:构建效用评价体系;研究无先验的用户社会关系的方法;研究大数据环境下的账号关联技术;跨社交网络的数据融合环境下的关联账号挖掘方法等.

6 结束语

各种功能社交网络的兴起,丰富了用户的日常生活,相应地产生了大规模用户信息. 但零散在多个平台的信息均不能全面刻画用户画像,成为身份关联技术发展的契机. 在此基础上,围绕着身份关联产生的用户大数据,依托深度学习等人工智能技术,实现人物画像的准确刻画,为发现社交网络舆情、分析用户行为特征及检查虚假消息等应用奠定了数据基础,进而能够为群体和个人决策提供有效帮助.

参考文献:

[1] Zhou X, Liang X, Zhang H, et al. Cross-platform identification of anonymous identical users in multiple social media networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(2): 411-424.

[2] Liu S, Wang S, Zhu F. Structured learning from heterogeneous behavior for social identity linkage[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(7): 2005-2019.

[3] Liu S, Wang S, Zhu F, et al. Hydra: large-scale social identity linkage via heterogeneous behavior modeling[C]// Proceedings of the 2014 ACM SIGMOD international conference on Management of Data. Snowbird: ACM Press, 2014: 51-62.

[4] Zafarani R, Liu H. Connecting users across social media sites: a behavioral-modeling approach[C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago: ACM Press, 2013: 41-49.

[5] Zhang Y, Tang J, Yang Z, et al. Cosnet: connecting heterogeneous social networks with local and global consistency[C]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney: ACM Press, 2015: 1485-1494.

[6] Korula N, Lattanzi S. An efficient reconciliation algorithm for social networks[J]. Proceedings of the VLDB Endowment, 2014, 7(5): 377-388.

[7] Lu C T, Shuai H H, Yu P S. Identifying your customers in social networks[C]// Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. Shanghai: ACM Press, 2014: 391-400.

[8] Kong X, Zhang J, Yu P S. Inferring anchor links across multiple heterogeneous social networks[C]// Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. San Francisco: ACM Press, 2013: 179-188.

[9] Goga O, Lei H, Parthasarathi S H K, et al. Exploiting innocuous activity for correlating users across sites[C]// Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro: ACM Press, 2013: 447-458.

[10] 叶娜, 赵银亮, 边根庆, 等. 模式无关的社交网络用户识别算法[J]. 西安交通大学学报, 2013, 47(12): 19-25.
Ye Na, Zhao Yinliang, Bian Genqing, et al. A schema-independent user identification algorithm in social networks[J]. Journal of Xi'an Jiaotong University, 2013, 47(12): 19-25.

[11] 徐钦. 基于遗传算法的复杂网络节点匹配问题[J]. 黑龙江科技学院学报, 2011, 21(3): 244-248.
Xu Q. Node matching between complex networks based on genetic algorithm[J]. Journal of Heilongjiang Institute of Science and Technology, 2011, 21(3): 244-248.

[12] 周小平, 梁循, 赵吉超, 等. 面向社会网络融合的关联用户挖掘方法综述[J]. 软件学报, 2017, 28(6): 1565-1583.

Zhou X P, Liang X, Zhao J C, et al. Correlating user mining methods for social network integration: a survey [J]. Journal of Software, 2017, 28(6): 1565-1583.

[13] 李晓菲, 梁循, 周小平, 等. 一种跨平台微博社区账户匹配方法[P]. 北京: CN104765729A, 2015-07-08.

[14] De Meo P, Ferrara E, Fiumara G, et al. On facebook, most ties are weak[J]. Communications of the ACM, 2014, 57(11): 78-84.

[15] Cheng Z, Caverlee J, Lee K, et al. Exploring millions of footprints in location sharing services[J]. ICWSM, 2011, 2011: 81-88.

[16] Zhang Y, Zhang M, Liu Y, et al. Localized matrix factorization for recommendation based on matrix block diagonal forms[C]//Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro: ACM Press, 2013: 1511-1520.

[17] Zhang Y, Zhang M, Zhang Y, et al. Daily-aware personalized recommendation based on feature-level time series analysis[C]//Proceedings of the 24th International Conference on World Wide Web. Florence: ACM Press, 2015: 1373-1383.

[18] ERDdS P, R&WI A. On random graphs I[J]. Publ Math Debrecen, 1959, 6: 290-297.

[19] Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks[J]. Nature, 1998, 393(6684): 440-442.

[20] Barabási A L, Albert R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509-512.

[21] Chakrabarti D, Zhan Y, Faloutsos C. R-mat: a recursive model for graph mining[C]//Proceedings of the 2004 SIAM International Conference on Data Mining. Lake Buena Vista: Society for Industrial and Applied Mathematics Press, 2004: 442-446.

[22] Man T, Shen H, Liu S, et al. Predict anchor links across social networks via an embedding approach[C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York: Morgan Kaufmann Press, 2016, 16: 1823-1829.

[23] Liu L, Cheung W K, Li X, et al. Aligning users across social networks using network embedding[C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York: Morgan Kaufmann Press, 2016: 1774-1780.

[24] Zhou X, Liang X, Du X, et al. Structure based user identification across social networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(6): 1178-1191.

[25] Zhou X, Liang X, Zhao J, et al. An unsupervised user identification algorithm using network embedding and scalable nearest neighbour [J]. Cluster Computing, 2018(3): 1-11.

[26] Bartunov S, Korshunov A, Park S T, et al. Joint link-attribute user identity resolution in online social networks [C]//Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis. Beijing: ACM Press, 2012: 1-9.

[27] Deng Z, Sang J, Xu C. Personalized video recommendation based on cross-platform user modeling[C]//2013 IEEE International Conference on Multimedia and Expo (ICME). San Jose: IEEE Press, 2013: 1-6.

[28] Mu X, Zhu F, Lim E P, et al. User identity linkage by latent user space modelling [C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM Press, 2016: 1775-1784.

[29] Acquisti A, Gross R, Stutzman F. Faces of facebook: privacy in the age of augmented reality[J]. BlackHat USA, 2011 (2): 1-20.

[30] Iofciu T, Fankhauser P, Abel F, et al. Identifying users across social tagging systems[C] //Fifth International AAAI Conference on Weblogs and Social Media. Barcelona: AAAI Press, 2011: 522-525.

[31] Motoyama M, Varghese G. I seek you: searching and matching individuals in social networks[C]//Proceedings of the Eleventh International Workshop on Web Information and Data Management. Hong Kong: ACM Press, 2009: 67-75.

[32] Zheng R, Li J, Chen H, et al. A framework for authorship identification of online messages: writing - style features and classification techniques[J]. Journal of the Association for Information Science and Technology, 2006, 57(3): 378-393.

[33] Al Mishari M, Tsudik G. Exploring linkability of user reviews[C]//17th European Symposium on Research in Computer Security. Pisa: Springer Press, 2012: 307-324.

[34] Nilizadeh S, Kapadia A, Ahn Y Y. Community-enhanced de-anonymization of online social networks[C]//Proceedings of the 2014 acm Sigsac Conference on Computer and Communications Security. Scottsdale: ACM Press, 2014: 537-548.

[35] Singh R, Xu J, Berger B. Global alignment of multiple

protein interaction networks with application to functional orthology detection[J]. Proceedings of the National Academy of Sciences, 2008, 105(35): 12763-12768.

[36] Pedarsani P, Figueiredo D R, Grossglauser M. A bayesian method for matching two similar graphs without seeds[C]//2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton). Monticello: IEEE Press, 2013: 1598-1607.

[37] Yu M. Entity linking on graph data[C]//Proceedings of the 23rd International Conference on World Wide Web. Seoul: ACM Press, 2014: 21-26.

[38] 孙昊. 大数据技术下的网络舆情分析系统研究[J]. 自动化与仪器仪表, 2018(8): 26-28.
Sun H. Research on network public opinion analysis system based on big data technology[J]. Automation & Instrumentation, 2018(8): 26-28.

[39] 余辉, 黄永峰. 跨媒体多源网络舆情分析系统设计与实现[J]. 数据通信, 2014(1): 39-42.
Yu H, Huang Y F. Design and implementation of cross-media multi-source network public opinion analysis system[J]. Data Communications, 2014(1): 39-42.

[40] 孟祥武, 纪威宇, 张玉洁. 大数据环境下的推荐系统[J]. 北京邮电大学学报, 2015, 38(2): 1-15.
Meng X W, Ji W Y, Zhang Y J. A survey of recommendation systems in big data[J]. Journal of Beijing University of Posts and Telecommunications, 2015, 38(2): 1-15.

[41] Le Q V, Mikolov T. Distributed representations of sentences and documents[C]//Proceedings of the 31st International Conference on Machine Learning. Beijing: ACM Press, 2014: 1188-1196.

[42] Chiu J P C, Nichols E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357-370.

[43] Tu C, Liu Z, Sun M. Inferring correspondences from multiple sources for microblog user tags[C]//Chinese National Conference on Social Media Processing. Beijing: Springer Press, 2014: 1-12.

[44] Mukherjee A, Venkataraman V, Liu B, et al. What yelp fake review filter might be doing? [C]//Seventh International AAAI Conference on Weblogs and Social Media. Massachusetts: AAAI Press, 2013: 409-418.

[45] Li J, Ott M, Cardie C, et al. Towards a general rule for identifying deceptive opinion spam[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Baltimore: ACL Press, 2014, 1: 1566-1576.

[46] Jindal N, Liu B. Opinion spam and analysis[C]//Proceedings of the 2008 International Conference on Web Search and Data Mining. Palo Alto: ACM Press, 2008: 219-230.

[47] Custard M, Sumner T. Using machine learning to support quality judgments[EB/OL]. 2005[2018-12-15]. http://www.dlib.org/dlib/october05/custard/10custard.html.

[48] Fogg B J, Soohoo C, Danielson D R, et al. How do users evaluate the credibility of web sites?: a study with over 2 500 participants[C]//Proceedings of the 2003 Conference on Designing for User Experiences. San Francisco: ACM Press, 2003: 1-15.