基于内容的社交网络用户身份识别方法

张树森 梁 循 弭宝曈 赵吉超 周小平
(中国人民大学信息学院 北京 100872)

摘 要 社交网络中识别用户身份具有重要价值,它对社交网络的分析与监管、用户行为的预测以及用户之间交互过程的研究具有重要意义.该文针对社交网络中的用户身份进行研究,将用户身份分为组织用户和个人用户,并对这两种用户身份进行具体定义和识别.该文研究问题属于社交网络用户分析研究中的子研究问题,主要通过用户在社交网络中发表的文本内容、多媒体内容以及用户时间序列内容识别出该用户的组织-个人身份,为社交网络用户身份的识别及进一步研究提供借鉴和帮助.在识别过程中,通过对文本内容中用户的口语化水平、内容(主题)复杂化水平、内容规范化水平的度量以及多媒体内容中用户图片特性和用户时间序列内容的分析,从不同角度提出 5 种机器可操作的用户组织-个人身份识别方法,进而识别出社交网络中用户是组织用户还是个人用户.最后,为了验证该文所提识别方法的可行性和有效性,该文选择新浪微博数据进行实验,并通过概率模型识别方法进行了对比分析.同时,在验证过程中,使用多种指标对实验结果进行评价.实验结果表明,该文识别方法能够有效识别出用户的组织-个人身份,其中内容复杂特性识别方法、内容规范化识别方法以及时间序列内容识别方法的用户身份识别准确率超过 80%.

关键词 社交网络;身份识别;内容复杂化;内容规范化;时间序列
中图法分类号 TP399 **DOI 号** 10.11897/SP.J.1016.2019.01739

Content-Based Social Network User Identification Methods

ZHANG Shu-Sen LIANG Xun MI Bao-Tong ZHAO Ji-Chao ZHOU Xiao-Ping
(*School of Information, Renmin University of China, Beijing* 100872)

Abstract It is of great value in social networks to recognise user's identity, which can help analyze and understand social networks, predict user's behavior, study the supervision of social networks and interaction between users. In this paper, we take user's identity in social networks as the research object. In view of the text content, multimedia content and time series content user post, we study the user's identity in social networks, which is divided into organization user and individual user, and the two user identities are defined and identified concretely. The problem in this paper is a sub-research problem in social network user analysis, and it mainly identifies the user's organization or individual identity through text content, multimedia content and time-series content published by social network users, which provided reference and help for the identification of social network user identity and further study. In social life, everyone has an identity in a specific environment relative to other people or things. Due to the different nature of users, organization users and individual users will have different characteristics in terms of social content. In social networks, user identities are essential parts of their social activities. User identities may be either explicit or implicit, and are more or less expressed in user-published content. At the same time, more user characteristics are also reflected in user postings. Therefore, the identity of social network

收稿日期:2017-10-13;在线出版日期:2018-05-28.本课题得到国家自然科学基金(71531012,71271211,71601013)、北京市自然科学基金(4172032,4174087)、北京教委科技计划项目(SQKM201710016002)资助.**张树森**,博士研究生,主要研究方向为数据挖掘、社会计算.E-mail:zss2446@163.com.**梁 循**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为数据挖掘、商务智能、社会计算.E-mail:xliang@ruc.edu.cn.**弭宝曈**,博士研究生,主要研究方向为物联网、社会计算.**赵吉超**,硕士研究生,主要研究方向为社会计算、网络舆情传播.**周小平**,博士研究生,讲师,中国计算机学会(CCF)学生会员,主要研究方向为社会计算、数据挖掘.

users can be analyzed and identified through the user's content. There are many kinds of information that can reflect the user's identity in social networks, such as tag information, registration information, labels or personal statements. However, it is difficult for us to identify the authenticity or credibility of these information. The registration information and authentication information in different social platforms are also different, so it is difficult to find common information that is suitable for many social platforms. However, the content information published by users on the social platform is basically visible, such as, micro blogging, Twitter, which are relatively basic and core parts of the social platform. And, it is also relatively easy to obtain. Therefore, the method of identifying user organization-individual identity based on user content information has a certain wide adaptability. Through measuring the user's colloquialization, content (theme) complexity and the normalization of text content, and simultaneously considering the user's picture characteristic in multimedia content and the user's time series content, we propose 5 kinds of machine-operated organization-individual identity recognition (identification) methods from different angles to identify whether the user in social networks is an organization user or a individual user. Finally, in order to verify the feasibility and validity of these methods mentioned in this paper, we choose Sina micro blogging dataset to experiment and through the probability model identification method for comparative analysis. At the same time, a variety of indicators are used in the verification process to evaluate the experimental results. The experimental results show that these methods can effectively identify the organization-personal identity of users in social networks, among which the recognition results of content complexity identification method, content normalization identification method and time series content identification method are the most ideal and the accuracy rate is above 80%.

Keywords social networks; identification; complexity of content; normalization of content; time series

1 引 言

当前,由于网络通讯技术和多媒体技术的快速发展,网络用户已从原始单纯的信息消费者转变成为信息的制造者[1].同时,Facebook、Twitter、微博、博客以及其它社交网络[2]应用的出现,使人类真实世界在网络虚拟世界中继续延伸.此外,由于人与人交互活动的途径呈数字化趋势,对社交网络的研究也逐步从社会学、社会行为及社会关系研究转变为数据挖掘、网络数理统计及量化分析研究.

在社交网络用户身份识别研究中,一些研究人员往往通过用户名的分析直接进行识别.但在以用户名为主要识别依据的方法中,很多情况下虽然能够识别出该用户的类别或身份,同时也存在较多问题.如随着社会的发展,不断有新的词汇或名称产生,像"逼格"、"蓝瘦"、"猴赛雷"等,当这些新出现的词汇出现在用户名中时,通常难以直接识别其类别或身份.在用户名中,也经常会有英文和字母缩写,以及用户自创拼写等的存在.当出现这些情况时,也往往难以辨别其类别或身份,如 TODS(上海 TODS 官云微博)、CYCAN(组织,中国青年应对气候变化行动网络官方微博)以及保健业跑酷(个人,公司总经理)、创意引领生活(个人,公司董事长兼总经理孔令博)、陈蓉博客(组织,上海新娱乐频道节目官方微博)等,单纯依靠这些用户名也容易产生错误的判断.实际上,在许多社交平台中存在认证用户,不过认证用户所占用户比例并不高.如我国新浪微博注册用户超过 5 亿,而认证用户只有几百万(约 300 多万,约占 0.6%),大部分用户并未进行认证.同时,社交网络中的实名制更多的是在后台显示,网络中主要还是用户的昵称.其中,在线社交网络由于其开放性、虚拟性等特点,用户社交范围限制较小.如果换成真实姓名,将对用户的活跃性产生较大影响.此外,如果换成真实用户姓名,也失去了用户的个性化,昵称是有必要保留的.

在社交网络中,用户身份是其社会活动中必不可少的一部分,它可能或明确或隐含,以及或多或少地表现在用户发表的内容中.同时,更多的用户特性也反映在用户发表内容中.因此,社交网络用户的身

份可通过用户内容进行分析和识别. 通过分析用户在社交网络中的内容信息、行为(如点赞、转发等)特征,对用户进行分类及个体定位[3]在实际应用和科研中具有重要意义. 在海量用户数据环境下,通过计算机自动实时识别本文用户组织-个人身份,不仅方便分类监管,还提高了相关部门监管和快速反应水平. 同时,识别社交用户身份对用户个性化服务、新闻推荐和产品营销等商业活动均具有重要的实践意义. 在研究中,本文根据用户内容信息识别用户组织-个人身份的方法具有一定的广泛适用性. 社交网络中,能够反映用户身份的信息有多种,如标记信息、注册信息、标签或个人说明等,但我们难以认定其真实性或可信度. 此外,不同的社交平台注册信息和认证信息内容也不尽相同,难以找到适用于众多社交平台的共有信息. 在社交平台上,用户发表的内容信息基本是可见的,如微博、Twitter 等,是社交平台较为基本和核心的部分,也是较为容易获取的. 同时,社交内容具有信息量丰富、数量多的特点,可以从不同角度进行分析. 此外,本文识别用户身份方法为后续具体身份(如白领、学生、教师等)的识别提供了方法上的借鉴或研究基础.

本文从不同角度提出 5 种识别方法,包括口语人称识别方法(Colloquialism and Person Identification Method,CPIM)、内容复杂特性识别方法(Content Complexity Identification Method,CCIM)、内容规范化识别方法(Content Normalization Identification Method,CNIM)、多媒体内容识别方法(Multimedia Content Identification Method,MCIM)和时间序列内容识别方法(Time Series Content Identification Method,TSCIM). 最后,在新浪微博数据集中对这些方法进行了验证,并通过概率模型方法进行对比分析. 实验结果表明,本文识别方法可有效识别出用户的组织-个人身份.

2　相关工作

2.1　现有工作

对社交网络用户的研究方法,通常是根据社交网络特征、内容信息以及用户动态行为表现对用户进行分析. 如在机器学习方法中,为识别用户角色,根据网络结构、交互内容等信息,通过无监督的学习过程,将用户自动分到不同角色类中. 社交用户身份的研究则通常是对用户的兴趣爱好、位置和社交关系等进行分析,推测不同用户所属群体或身份[3],以及对用户身份同一性判定[4]和用户身份认证[5]方面的研究. 其中,用户身份同一性的识别方法主要集中在用户各种属性字段的匹配,即各种属性相似度计算. 如 Raad 等人[6]提出一种 FOAF(Friend-Of-A-Friend)属性匹配的方法,将用户属性资料提取转化为统一的 FOAF 格式,得到属性值之间的相似度. Cortis 等人[7]提出一种带有权重的基于用户资料的身份识别算法,从语法和语义两个方面计算属性资料的相似度等. 为了更好的进行身份同一性识别,而不是单纯依靠属性匹配,研究人员又加入好友链接关系和发表内容等识别信息. 如 Kong 等人[8]提出的 MNA(Multi-Network Anchoring)算法. 还有一些研究者仅从用户名角度出发进行用户身份同一性的识别,如 Liu 等人[4]提出一种基于用户名特征的用户身份同一性判定方法. 在用户身份认证方法中,根据生物特征识别越来越得到重视,如基于虹膜[9]、步态[10]的身份识别认证,以及常见的基于掌纹(或指纹)、人脸等. 社交用户身份的认证主要依据其他认证因素. 如 Brainard 等人[11]提出将"你认识的人"(Somebody you know)作为认证因素,即依赖社会关系进行识别.

社交网络中与用户身份相关的研究主要还有从社交网络中推断用户的各种属性,包括性别、年龄、职业、位置、兴趣、政治倾向、经济状况等. 如 Yang 等人[12]发现个人兴趣和友情信息是高度相关的,提出了一个朋友-兴趣传播模型(Friendship-Interest Propagation model,FIP model),从而得到用户友情和兴趣的预测情况. Luo 等人[13]根据用户在社交网络中所处地位分析出该用户的经济状况等.

在社交网络内容方面,如社交用户情感分析、信息传播机制、事件检测等研究中,社交内容均具有重要地位. 如在社交网络内容分析中,Subbian 等人[14]通过分析社交网络内容的主题信息并结合时间敏感性的方法得出用户实时影响力. 文献[15-17]通过情感词方法对社交网络文本内容进行分析,实现用户情感分析. 在对社交网络中图片、视频等多媒体内容的处理中,You 等人[18]利用微调的深度卷积神经网络构架训练图片情感分析模型,并取得较好效果. Chao 等人[19]利用长短期记忆神经循环网络构架和时间池技术,实现对音频和视频内容情感分析. 此外,Li 等人[20]利用短文本数据集和时间信息,提出微博突发事件监测增量时间主题模型,实现微博突发事件的监测. Wang 等人[21]基于非参数贝叶斯理论提出一种融合内容信息和时间序列信息的潜在影响力传播模型 InfoIBP(Influence propagation on Indian Buffet Process),实现对用户潜在影响力的

研究和分析.

此外,对社交网络用户及内容进行分析面临着众多的挑战和困难,如社会网络复杂性分析问题、海量数据处理问题、数据动态变化问题、数据异构问题、评估机制问题、数据采集问题以及个人隐私安全问题等,需要我们不断研究和解决.

2.2 社交内容表示及相关定义

在本文中,我们用集合 $A=\{V,C\}$ 表示社交网络用户的社交内容.其中,V 表示 N 个用户的集合,即 $V=\bigcup_{i=1}^{N} v_i$,C 表示 N 个用户内容集合,$C=\bigcup_{u\in V} C_u$,C_u 为用户 u 的内容集合,$C_u=\{X_u,M_u,S_u\}$,X_u,M_u,S_u 分别为用户 u 的文本内容集合,多媒体内容集合,及其中的时间序列内容集合.

(1) $X_u=\{x_1,x_2,x_3,\cdots,x_n\}$,$x_i$ 为用户 u 发表的第 i 条文本内容,$|X_u|$ 为该用户发表文本内容条数 n,$|x_i|$ 为第 i 条文本内容长度.

(2) $M_u=\{p_1,p_2,p_3,\cdots,p_n\}$,$p_i$ 为用户 u 发表的第 i 张图片,$|M_u|$ 为该用户发表图片数目 n.

(3) $S_u=\{x_1=\langle t_1,P(t_1)\rangle,x_2=\langle t_2,P(t_2)\rangle,x_3=\langle t_3,P(t_3)\rangle,\cdots,x_n=\langle t_n,P(t_n)\rangle\}$,其中,$t_i$ 为用户发表内容时间 i 并按先后顺序递增,$i\in[1,n]$,$P(t_i)$ 表示在 t_i 到 t_{i+1} 时间段内 $[t_i,t_{i+1})$,用户 u 发表内容数目占所有内容数目的比例,其表达式为

$$P(t_i)=\frac{f(t_{i+1}-t_i)}{n} \tag{1}$$

$f(t_{i+1}-t_i)$ 为用户 u 在时间段 $[t_i,t_{i+1})$ 内发表内容数目,$P(t_i)\in[0,1]$,$i\in[1,n]$.

定义 1. 个人用户(Individual) v_{Ind},本文指的是社交网络虚拟空间中的某一节点,该节点所代表的用户基本属性是有意识或有较高层次的单独的人类(Pe)个体,$v_{\mathrm{Ind}}\in Pe$.

定义 2. 组织用户(Organization) v_{Org},本文指的是社交网络虚拟空间中的某一节点,该节点所代表的用户是按照一定的目的、任务和形式编制起来的社会集体或团体,$v_{\mathrm{Org}}\notin Pe$.

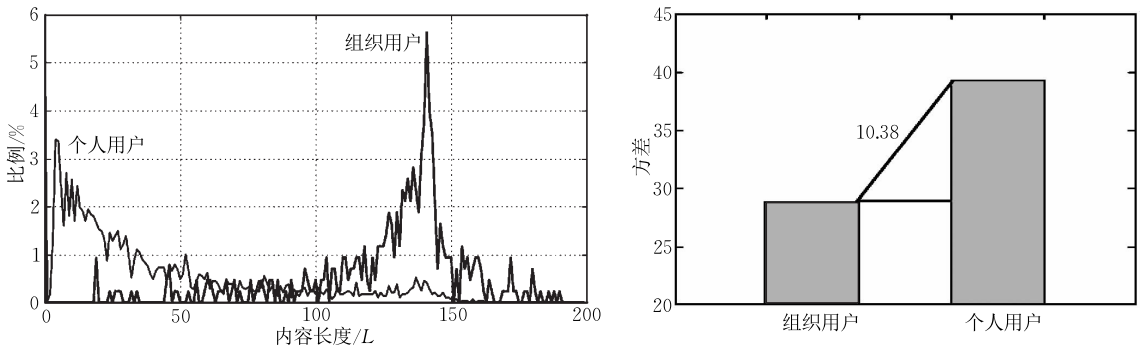
在社交网络中,用户通常随机发表各种言论,这些言论随着用户的状态、时间、环境以及心情的不同而不同,难以找到一个统一的主题.在内容主题方面显得十分混乱、复杂.同时,用户发表的内容在长度及格式规范上通常会呈现出长短不一,格式多样,具有不同标记、符号或者表情等特征,使得内容结构上同样显得十分混乱、复杂.此外,用户发表的图片内容通常呈现出多元化、同质性小的特点,且其中出现人物通常变化较快、表情不一,使得多媒体内容也显得十分混乱、复杂.

上述中出现的"混乱"一词,本文定义了其具体含义.

定义 3. 混乱(Confusion)指的是社交网络中,用户发表内容所表现出来的无序性,没有条理性或者差异变化较大的特性.

在社交网络中,个人用户和组织用户通常根据自身需要发表各种内容.由于自身属性的不同,发表的内容会表现出不同的内容特征.个人用户发表的内容往往与用户自身的心情、兴趣以及所在环境相关,较多的表现出人类共有的基本属性.而组织用户由于自身组织属性,发表的内容往往具有明显的宣传、通知、倡议以及信息发布的特性,并且往往能够表现出明显的主题和目的.因此,从发表内容的主体和结构方面来讲,个人用户混乱程度(C)通常要高于组织用户,即 $C(v_{\mathrm{Ind}})>C(v_{\mathrm{Org}})$.

在实际研究中,社交内容长度的分布情况可以反映出不同身份用户内容的整体混乱程度,能够从宏观角度反映出不同身份用户内容的整体结构状况.通过对两种身份用户数据的分析和计算,本文得出两种身份用户(统计部分两种身份用户数据)文本内容平均长度分布及方差,如图 1 所示.



(a) 内容长度 (b) 方差

图 1 用户内容长度分布、方差比较图

在图 1(a)用户内容长度分布比较图中，内容长度为 0 表示用户发表的内容中没有文本内容，只有符号、表情或图片等内容. 在本文所用数据集中，通过对所有用户发表的微博内容长度分析，可以得出个人身份用户平均方差约为 39.28，组织身份用户平均方差约为 28.9. 由此可以看出，个人身份用户的内容长度波动性相对更大，从用户内容长度角度来讲，其整体混乱程度也更高.

3　用户身份识别方法

在社交网络中，用户发表的内容，如中文微博，通常具有语言简短灵活、文本不规范、噪声较大等特点[22]. 同时，社交网络用户发表的文本内容及其语法结构通常不规范，用语也不规范，甚至许多是用户自定义的标记或符号[23-24]. 本文针对以上特性及对社交内容的分析，从多个角度提出本文用户组织-个人身份识别方法，为社交网络用户身份的进一步研究提供借鉴和帮助.

3.1　文本内容分析识别方法

3.1.1　口语人称识别方法(CPIM)

在社交网络中，由于用户自身属性，不同身份用户发表的内容在口语使用和人称方面会有一定差别. 本文中所说的口语是区别于“书面语”的，人们日常口头交谈时说的通俗语言，如“哎呀”、“哥儿们”、“着调”等. 为了能够有效识别出用户是个人身份还是组织身份，本文通过度量用户内容的口语化比率，以及第一人称使用的人称比率对用户发表的文本内容进行量化分析，进而为两种身份的识别提供识别依据.

定义 4.　在社交网络中，本文基于用户 u 发表文本内容 $X_u=\{x_1,x_2,x_3,\cdots,x_n\}$，定义用户口语化比率(Colloquialization ratio) v_c 及人称比率 v_p(Person rate)为

$$v_c=\frac{\sum_{i=1}^{i=n}f(x_{ic})}{n} \tag{2}$$

$$v_p=\frac{\sum_{i=1}^{i=n}f(x_{ip})}{n} \tag{3}$$

其中 $f(x_{ic})$ 表示内容 x_i 中口语数函数，$f(x_{ip})$ 表示内容 x_i 中第一人称数函数.

在社交网络中，个人身份用户在发表内容时，往往更倾向于随意发表自己的看法，表达自己的感情，其内容一般表现出通俗易懂，表达直接，口语化相对明显的特点. 而组织用户则由于其自身团体属性的限制和影响，其发表的内容更多的是一些书面语言，表达更为理性，口语使用相对较少. 通过对不同身份用户口语化和人称的度量，实现从微观角度对不同用户内容的分析和识别.

3.1.2　内容复杂特性识别方法(CCIM)

在社交网络中，由于用户所处环境及自身属性的不同，个人用户和组织用户发表内容通常表现出不同的内容复杂特性. 本文通过对内容主题多样性的度量，实现这种复杂特性的量化分析. 根据信息熵的原理，如果用户发表各个内容之间的整体相似度越小，则该用户内容不确定性越大，其所含信息量就会越多，即认为该用户所发内容包含主题会越多，内容也越复杂. 本文提出了用户内容熵的概念，用来度量用户 u 的内容复杂特性水平.

定义 5.　基于用户 u 发表的文本内容 $X_u=\{x_1,x_2,x_3,\cdots,x_n\}$，本文将用户的内容熵值 E_u 定义为

$$E_u=-\frac{\sum_{i=1}^{n}\sum_{j\in x}^{j\neq i}[(1-S_{ij})\log_2 S_{ij}]}{n(n-1)} \tag{4}$$

其中，S_{ij} 为用户发表的第 i 条内容 x_i 和第 j 条内容 x_j 之间的相似度，且 $S_{ij}\in(0,1)$.

在文本处理研究中，文本的相似性度量方法有多种，可使用最长公共子序列(Longest Common Subsequence，LCS)、编辑距离、语言模型、主题模型以及余弦相似度等，实现对文本内容间相似度的度量或计算.

通常来讲，个人用户内容主题较多，内容较丰富，其内容复杂特性相对较高，内容间的相似度 S 较小，其对应的熵值 E_u 也大. 组织身份用户在社交网络中发表的内容则通常与组织的性质有关，其主题会更加突出，因而其内容复杂特性相对较低. 本文通过对用户内容熵的计算，实现用户内容复杂特性的度量，从而将用户整体的内容复杂特性抽象或刻画出来. 由此，从用户内容的宏观角度，为用户组织-个人身份识别提供识别的依据.

3.1.3　内容规范化识别方法(CNIM)

在社交网络中，组织用户通常在内容的格式方面具有一定的要求，内容语句结构也更加符合规范，错误较少等. 而个人用户在发表内容的格式规范方面通常没有组织用户要求高，内容格式整体上显得更加混乱，内容长度总体波动相对来说较大，很少具有统一的格式. 由此，本文提出结构熵的概念来度量这种混乱程度，用户内容体系结构越规则，熵就越小，越不规则熵值就越大.

本文根据用户发表内容整体长度情况和用户内容格式的实际情况度量用户内容的规范化水平.其中,在考虑用户内容格式过程中,当用户发表的内容具有统一的格式规范时,如发表内容开头都在中括号内表明主题,许多发表的内容含有其他相同标志或统一的符号等,本文将其称为含有统一标志的内容.

定义 6. 基于用户 u 发表的文本内容 $X_u=\{x_1,x_2,x_3,\cdots,x_n\}$,我们令,

$$f(x_i)=\frac{|lx_i-\mu|}{\mu} \tag{5}$$

$$h=\frac{1}{n}\sum_{i=1}^{i=n}[1-f(x_i)] \tag{6}$$

$$E_u=-\left(\frac{N-m_x+1}{N}\right)\log_2 h \tag{7}$$

其中,lx_i表示内容 x_i的长度,μ 为用户内容平均长度;$f(x_i)$表示用户发表内容 x_i的"变异系数",即 x_i离散程度的测度值;h 表示用户内容 X_u中 $1-f(x_i)$的均值,可看作是用户内容的集中程度(相对离散程度来讲,离散程度越高,则集程度越低),h 值越大集中程度越高;m_x为含有统一标志的内容个数,N为用户发表内容总数,E_u为用户内容结构熵值.

在式(7)中,不考虑内容格式,如果一个用户发表内容结构(长度)上越不规则,整体混乱程度越大,则 $f(x_i)$值就会越大,h 值就会越小,从而其熵值 E_u就会越大.当考虑内容格式时,用户内容格式越规范,熵值 E_u就越小.

由此,本文通过上述定义的结构熵值 E_u来识别用户的组织-个人身份.同样从用户内容的宏观角度,为识别该用户的组织-个人身份提供判别依据.

3.2 多媒体内容识别方法(MCIM)

在社交网络中,用户发表的内容除了文本内容之外,通常还会将自己喜欢的图片、音频以及视频等多媒体内容和文本内容一并发布出去,甚至单独发布出去.本文通过对两种不同身份用户所发图片中包含的人物进行分析,对用户进行识别.

在社交网络中,通过分析我们可以发现,个人用户的图片中同一人会更多的在多张图片中出现.组织用户的图片中人物差异较大,能够在多张图片中共同出现的同一人数目较少,图片的内容更多的是与该组织性质相关.由此,本文将用户身份识别问题转换为图片中人物相似性分析问题.依据图片中人物的相似性度量,在社交平台中对用户发表的图片内容进行人脸识别,获取图片中人脸的特征信息,对多张图片中出现的人物进行相似性的比较,判断是否为同一人.若为同一人的图片数量大于设定的一个阈值,则认为是个人身份,否则为组织身份.在人脸识别的方法中,常见的有基于 PCA、基于 Fisher 线性判别以及基于 LBP 特征的人脸识别算法等.在具体实验中,本文所采用的是基于 PCA 的人脸识别算法.

在得到用户图片信息识别结果后,本文通过定义图片比率 P_u,即用户发表图片中包含同一人物图片数目与该用户所有图片数目的比值,来识别用户的组织-个人身份.基于用户 u 发表图片内容:$Mu=\{p_1,p_2,p_3,\cdots,p_n\}$,图片比率 P_u表示为

$$P_u=\frac{\sum_{i=1}^{i=n}f(p_i)}{n} \tag{8}$$

其中 $f(p_i)$表示图片 p_i中是否为含有同一人的函数并计算其函数值,n 为用户图片总数.

在社交网络中,本文组织-个人身份用户图片内容的处理、识别过程主要包括以下几个步骤:

步骤 1. 图片预处理,获取用户发表的图片并对图片进行统一、规范化处理.

步骤 2. 构建特征脸空间,确定人脸特征并将用户待识别图片投影到特征子空间.通过投影向量比较,得出图片人脸识别结果.

步骤 3. 人物相似性判断,根据用户图片人脸识别结果分析在多张出现同一人的最大数目,得到该用户同一人的图片比率.

步骤 4. 获得用户身份识别结果,根据设定的图片比率识别阈值 $IT(P_u)$,若该用户大于该阈值,则认为该用户为个人用户,否则为组织用户.

3.3 时间序列内容识别方法(TSCIM)

3.3.1 用户社交行为时间序列

在数据挖掘中,时间序列(Time series,Ts)指的是按时间顺序获得的一系列观测值[25].时间序列不仅表达了数据随时间变化的规律,还表达了数据分布的时间规律.在社交网络中,不同用户的社交行为会随着时间的发展变化,呈现出一定的变化规律.其中,社交网络用户发布信息的行为也会随着时间的变化而变化,呈现出一个用户社交行为时间序列,即不同时间段内,用户发表信息的多少会有不同的变化.组织用户和个人用户信息发布数量的分布同样会随着时间变化表现出不同的变化规律.

通过对本文社交网络数据的处理和分析,针对用户在一天 24 小时不同时间段内发布信息的归纳,可发现以下 3 条规律:

(1) 个人用户和组织用户具有不同的信息发布时段和峰值.通常来讲,组织用户有两个较明显的信

息发布时段，即 9:00～11:00、15:00～17:00，且具有三个明显峰值，这些信息发布时段和主要(两个最高)峰值均处于工作时间. 而个人用户则可明显分为集中和稀疏两个数据部分，信息发布时段没有太大的变化，具有两个明显峰值，峰值均处于休息时间，即 11:00～13:00、21:00～23:00.

(2) 个人用户和组织用户的信息发布数量随着时间的变化呈现不同的变化趋势. 组织用户在夜间变化幅度较小，基本处于平稳状态，而在白天变化波动较大. 个人用户在一天内整体呈现平稳变化的趋势，夜间变化幅度较大，白天则波动较小.

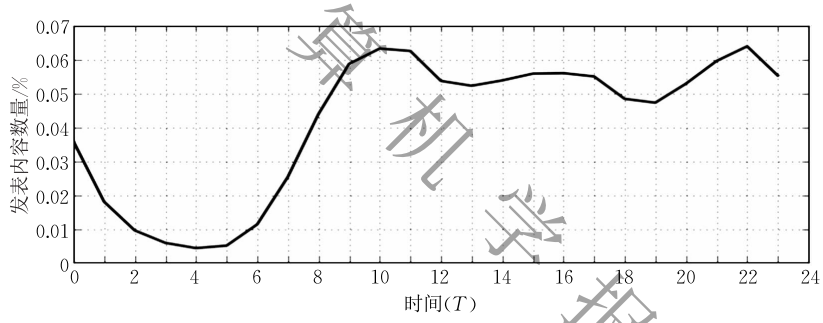
(3) 个人用户和组织用户的信息发布数量在不同时间段具有较大的差别. 由此可以表明，在各个时间段内组织用户和个人用户活跃程度具有明显差别.

由上述规律可以看出，组织用户和个人用户在信息发布的社交行为时间序列方面存在显著差异. 因此，本文通过对用户时间序列的分析，同样能够识别出该用户的组织-个人身份.

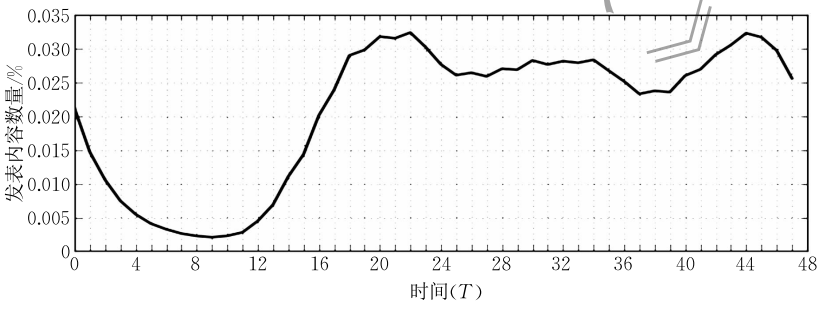
3.3.2　时间序列识别方法

为识别社交网络用户的组织-个人身份，本文提出了基于经验思路进行识别的方法. 通过对两种身份约 2 万用户数据的归纳得到两种身份用户的时间序列，并将其近似看作两种身份用户的标准时间序列. 以此为基础，通过计算社交网络用户与这两种标准时间序列的相近函数值，并根据这两个函数值的大小识别出该用户的组织-个人身份.

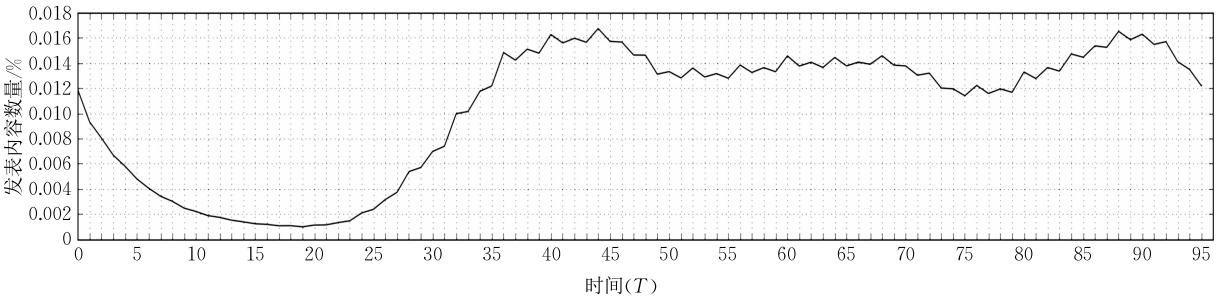
由于时间序列通常具有数据量大、更新快、噪音高、维数高的特点，因此，难以直接在原始时间序列上进行分析和归纳. 本文通过对时间序列分段线性表示，以固定窗口划分的方式对社交网络中的时间序列数据进行分析. 针对社交网络用户的时间序列数据，本文通过不同粒度(granularity，gr)大小划分时间序列关键点. 通过分析组织和个人用户数据，本文得出个人用户与组织用户标准时间序列如图 2、图 3 所示. 其中，时间序列图中随着粒度的不断增



(a) $gr=24$

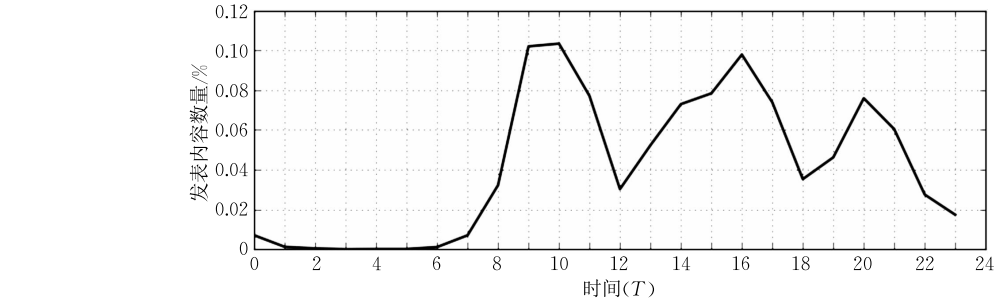


(b) $gr=48$

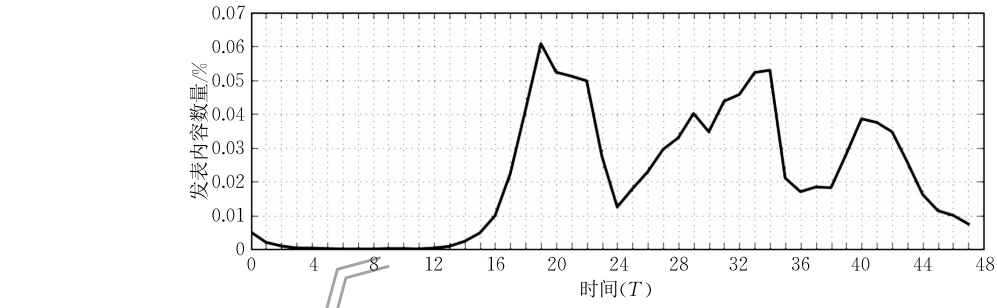


(c) $gr=96$

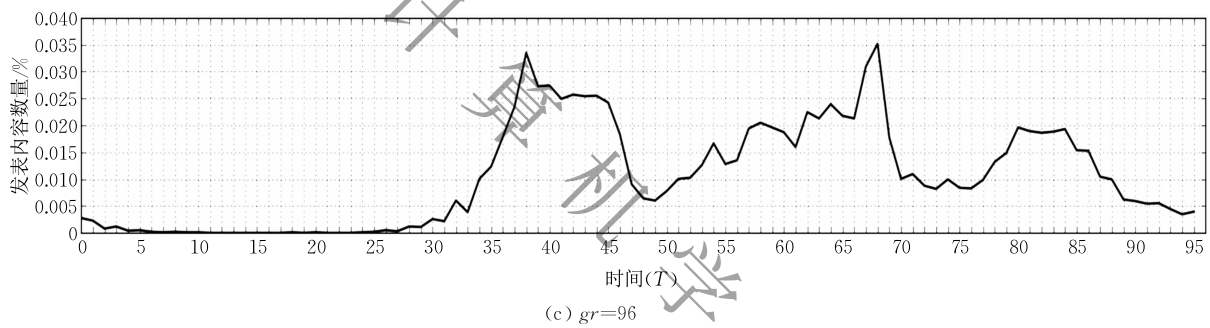
图 2　个人用户标准时间序列



(a) $gr=24$



(b) $gr=48$



(c) $gr=96$

图 3 组织用户标准时间序列

大,用户时间序列信息会越来越详细.本文通过与不同用户的时间序列进行对比分析,并计算两种身份用户相近函数值,从而识别出用户的组织-个人身份.

(1)用户标准时间序列及表示

本文基于用户 $u(u\in V)$的时间序列:$S_u=\{x_1=\langle t_1,P(t_1)\rangle,x_2=\langle t_2,P(t_2)\rangle,x_3=\langle t_3,P(t_3)\rangle,\cdots,x_n=\langle t_n,P(t_n)\rangle\}$,通过对时间序列图的分析可以发现时间序列随着时间变动,在两个关键点(如 t_i、t_j,两者之间时间序列只有一种变化趋势)之间一般会有三种变化趋势的可能,即-、\、/.这三种趋势分别代表时间序列值在该时间段内不变、下降、上升.在处理数据时,本文分表用 a,b,c 代表这三种趋势.由此,本文可通过这三个字母的一个组合序列,将用户的时间序列在不同时段的趋势简单的表示出来.实际上,这种时间序列表示方法无法反映时间序列幅度变化的大小,刻画的仅仅是两个关键点间的趋势.如图 4(示例)所示,其时间序列 $Ts[T_3,T_{12}]$可以表示为 $Ts=$'accccbbbb'.

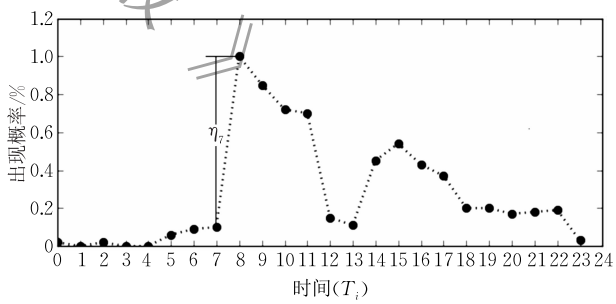


图 4 时间序列 Ts

同样,根据图 2、图 3 中用户标准时间序列,本文可得出两种身份用户在划分粒度分别为 $gr=24$、$gr=48$、$gr=96$ 下的标准字母序列表示.

(2)时间序列身份识别

在具体身份识别过程中,本文得出一个用户的时间序列后,应当同时考虑用户时间序列的趋势和幅度进行识别.由此,本文需要计算该用户时间序列与两种身份标准序列之间的相近(Close)函数 $F(c)$.此外,本文用户时间序列和标准时间序列均为齐序列.

本文将时间序列相近函数公式 $F(c)$ 定义为

$$F(c)=\mu\sum_{i\in S(match)}^{S(match)}(\eta_i/\sum\eta)\frac{L_{match}}{L_{ts}}+(1-\mu)\sum_{j\in S(LSS)}^{S(LSS)}(\eta_j/\sum\eta)\frac{L_{LSS}}{L_{match}} \tag{9}$$

其中，μ 为参数，用来控制不同变量值的比重；L_{match} 为对应位置字母序列匹配数目，L_{ts} 为字母序列总长度(或个数)，L_{LSS} 为最长子序列 LSS(连续)长度；$S(a)$ 表示 a 的匹配集合，即匹配的字母序列对应的位置号；η_i 表示时间 i 序列值与时间 $i+1$ 序列值的差值(变化的幅度)，$\sum\eta$ 表示时间序列总差值. 相似度这样定义是为了在匹配率相同的情况下，通过最长子序列比率(本文中为最长子序列与匹配序列长度比率)控制衡量相近函数值. 由此，本文能够根据社交网络用户的时间序列识别出该用户的组织-个人身份.

根据本文提出的用户组织-个人身份识别方法，本文进一步通过真实的社交网络数据进行验证，从而说明本文所提社交网络用户组织-个人身份识别方法的有效性和可行性.

4　实验及结果分析

在实验过程中，本文选择新浪微博数据进行验证. 通过对微博用户组织-个人身份的识别，验证本文所提识别方法的有效性和可行性. 同时，使用常见的概率模型识别方法进行对比.

4.1　实验数据及评价指标

(1) 实验数据

本文新浪微博实验数据集中包括个人身份用户 46 153 名(约占总数 71%)，组织用户 18 809 名(约占总数 29%)，共计 6.5 万余名微博用户. 其中，个人身份用户所发微博总数约为 2078 余万条，组织身份用户所发微博约为 1109 余万条，总计 3200 余万条微博. 本文实验数据中不同用户微博数量分布如图 5 所示.

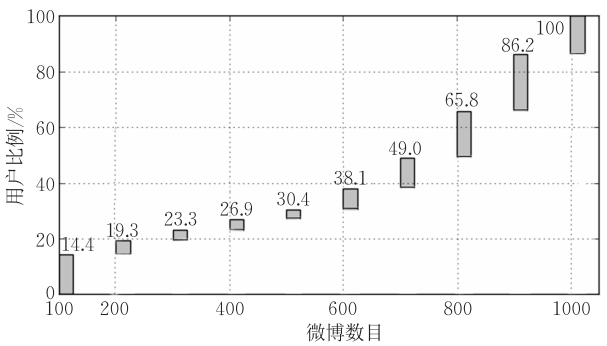


图 5　用户分布瀑布图

由于本文选择的微博用户已通过微博官方认证(如加 V 认证用户)，能够确定该用户是一个组织身份还是一个个人身份，进而为验证本文识别方法的有效性和可行性提供帮助.

(2) 评价指标

在实验过程中，本文采用常见的多种评价指标(Evaluation Indicators，EIs)对用户组织-个人身份识别结果进行评价，包括准确率、精确率、召回率、F1-Score. 在本文识别过程中，用户身份识别结果只有个人用户和组织用户两种结果(即或者识别为个人用户，或者识别为组织用户). 识别方法对两种身份的识别具有不同的精确率、召回率以及 F1-Score 指标. 本文在实验中分别对精确率、召回率和 F1-Score 指标进行了评价.

4.2　实验过程

在实验中，我们首先通过常见概率模型方法对用户组织-个人身份进行识别，并将该方法识别结果与本文识别方法进行对比.

4.2.1　概率模型识别(PM)

根据贝叶斯定理，建立微博用户身份识别概率模型(Probability Model，PM). 在识别过程中，以用户发表内容中包含的词为分析对象，通过贝叶斯定理的思想(朴素贝叶斯)，求解该用户微博内容特性条件下两种身份出现的概率，将该用户身份出现概率最大的作为该用户身份. 即根据微博内容所属用户身份的概率，得出该用户的身份.

其中，某一身份下，词概率公式 $P(S_j/Id_i)$ 为

$$P(S_j/Id_i)=\frac{NS_j+1}{MS_i+VS} \tag{10}$$

其中，NS_j 为身份 Id_i 中包含该词 S_j 的微博条数，VS 为所有微博中包含的词(无重复). 我们假设用户的一条微博为 wb，其内容提取的词为 $s_1, s_2, s_3, \cdots, s_n$，共有 n 个词，则该微博用户身份($Id_i, i=0,1$)计算概率公式 $P(Id_i/wb)$ 为

$$P(Id_i/wb)=P(s_1/Id_i)\times P(s_2/Id_i)\times P(s_3/Id_i),\cdots,P(s_n/Id_i)\times P(Id_i) \tag{11}$$

由此，通过求用户微博的身份概率，得出用户 user(假设该用户有 n 条微博 wb)的身份. 在概率模型识别方法下用户组织-个人身份识别结果如表 1 所示.

表 1　概率模型识别评价指标　(单位：%)

User	*Precision*	*Recall*	F1-Score	*Accuracy*
Ind	57.84	79.38	66.92	59.4
Org	63.22	37.98	47.46	

由表1可知,概率模型识别方法准确率为59.4%.由此可看出,通过概率模型的方法也可将用户的组织-个人身份近似识别出来.而识别率接近60%,也说明这两种身份用户的社交内容在用词方面并没有明显的区别.

4.2.2 口语人称识别(CPIM)

本实验中需要建立一个适用于社交网络的口语库,本文称之为社交口语库.我们将汉语词典中标记为口语的词组提取出来,组建一个初始口语库(不包括常用语).我们统计超过3万名微博用户发表内容的口语使用情况,并将初始口语库中使用数量低于5的口语词组从初始库中去掉,使初始口语库规模变小.最终得到更加适用于社交网络的社交口语库,其中口语的数目为1235个.

本文通过多次实验比较发现,个人用户口语化比率 sl 通常在0.2以上,而组织用户一般会在0.2以下.因此,本文在 $sl=0.2$ 时获得最大识别率(准确率)67.4%.同时,在用户人称使用方面,通过对实验发现个人身份用户第一人称比率 pe 通常在0.6以上,而组织身份用户则通常在0.6以下.因此,用户内容第一人称比率 $pe=0.6$ 时获得最大识别率61%.

本文综合考虑口语和人称识别依据,将识别阈值 $IT(SP)=(sl,pe)$ 设置为 $IT(SP)=(0.2,0.6)$,并将其作为用户口语人称识别模糊界限,其识别率为53.64%.本文不同阈值实验结果,如表2所示.根据表2识别结果,口语人称识别中不同阈值下的用户身份识别准确率,如图6所示.

表2 口语人称实验准确率 (单位:%)

识别依据	阈值(IT)-sl、pe									$IT(SP)$ (0.2,0.6)
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
口语	52.8	67.4	54.0	40.0	25.4	20.3	23.5	23.0	22.1	53.64
人称	26.3	29.7	30.5	32.0	44.0	61.0	51.8	45.8	32.0	

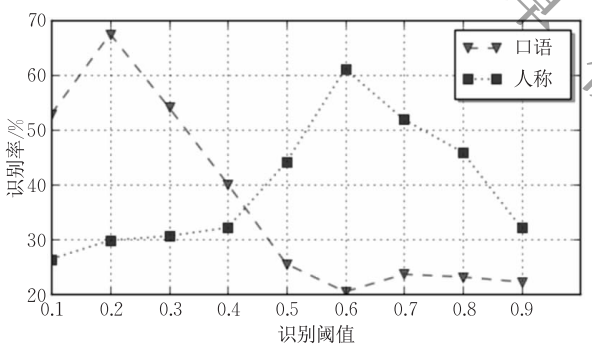


图6 CPIM准确率图

由图6可知,随着识别阈值的增加个人用户和组织用户识别准确率差别较大.个人用户口语化比率通常比组织用户高,当阈值 sl 低于0.2时,大部分个人用户会被识别出来,且部分组织用户会被识别为个人用户,准确率在阈值 $sl=0.2$ 时取得最大值67.4%.当阈值 sl 大于0.2时,随着阈值的增大,越来越多的个人用户将被识别为组织用户,因此总体呈现为下降的趋势.同样在人称方面,个人身份用户内容中第一人称比率同样大于组织身份用户,准确率在阈值 $pe=0.6$ 是取得最大值61%,当阈值 pe 小于0.6时,大部分个人用户会被识别出来,而随着阈值减小,越来越多的组织用户会被识别为个人用户.当阈值大于0.6时,随着阈值的增大,越来越多的个人用户将被识别为组织用户,因此总体呈现为下降的趋势.同时将二者考虑时,其识别率仅为53.64%.该方法实验效果并不理想,其主要原因是这两种识别因素对这两种身份用户的刻画存在共用问题,以及识别要求的提高(同时考虑两种识别因素).此外,该方法也为我们了解这两种身份用户的语言特征和差别提供了帮助.由于识别率较低,其他评价指标价值不大,不再详细描述.实际应用中我们可以根据需要,选择口语人称识别及综合其中各个识别结果进行最终的判别.

4.2.3 内容(主题)复杂特性识别(CCIM)

根据内容熵计算公式(4),我们可以计算出用户内容熵值.实验中,本文选择LCS的方法来计算相似度S.LCS是得到两个给定序列 s_i,s_j 的子序列 lcs,该子序列在两个序列中以相同的顺序出现,但不必要是连续的.实验中,本文内容相似度S定义为

$$S_{ij}=\text{len}(lcs)/\text{len}(\max(s_i,s_j)) \tag{12}$$

其中,len(a)表示计算对象 a 的长度函数,max(s_i,s_j)表示两个给定序列 s_i,s_j 中长度最大序列.

为了识别用户的组织-个人身份,本文设置不同内容熵值作为用户身份的识别阈值,识别结果如表3所示(内容熵值:E_u,个人用户:Ind,组织用户:Org).

实验过程中,对社交网络用户身份的识别结果有两种可能,要么识别为组织用户,要么识别为个人用户.本文不同阈值下识别准确率,如图7所示.

表 3　CCIM 准确率　（单位：%）

User	$IT(C)$-E_u							
	1.5		2		2.5		3	
	Ind	Org	Ind	Org	Ind	Org	Ind	Org
Ind	99.9	0.1	98.3	1.7	93.3	6.7	78.3	21.7
Org	91.3	8.7	48.6	51.4	21.2	78.8	2.1	97.9
Accuracy	73.49		84.72		89.10		83.97	

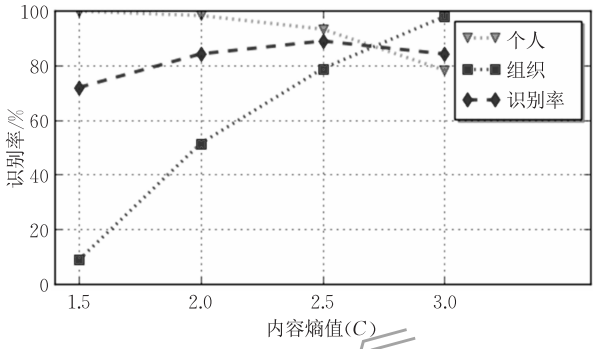


图 7　CCIM 准确率

由图 7 可以看出，随着阈值的增加个人身份用户准确率呈下降趋势，而组织身份用户呈上升趋势. 在设置的识别阈值中，当识别阈值 $IT(C)$ 为 2.5 时，本文取得最大识别准确率 89.1%. 同时，本文计算其他评价指标(EIs)如表 4 所示.

表 4　CCIM 评价指标(EIs)　（单位：%）

EIs	$IT(C)$-E_u							
	1.5		2		2.5		3	
	Ind	Org	Ind	Org	Ind	Org	Ind	Org
Precision	99.9	8.7	98.3	51.4	93.3	78.8	78.3	97.9
Recall	72.9	97.3	83.2	92.5	91.5	82.7	98.9	64.8
*F*1-Score	84.3	15.9	90.1	66.1	92.4	80.7	87.4	77.9

由表 4 可以看出，随着识别阈值增加，个人用户召回率呈上升趋势，而组织用户呈下降趋势. 当识别阈值为 2.5 时，用户内容复杂特性识别方法中 *F*1-Score 取得最佳值，即该方法识别个人用户时 *F*1-Score 为 92.4%，识别组织用户时 *F*1-Score 为 80.7%.

4.2.4　内容规范化识别(CNIM)

本实验中我们计算用户内容的结构熵值来判断该用户的组织-个人身份. 通过对比用户发表所有内容中是否出现共有的格式或规范，确定该用户是否含有统一标志的内容及个数. 本文设置不同结构熵值作为组织-个人身份的识别阈值 $IT(S)$，识别准确率如表 5 所示(结构熵值：E_u). 实验过程中，根据实验结果得到在不同识别阈值下，用户组织-个人身份识别准确率，如图 8 所示.

表 5　CNIM 准确率　（单位：%）

User	$IT(S)$-E_u							
	0.8		1.0		1.2		1.4	
	Ind	Org	Ind	Org	Ind	Org	Ind	Org
Ind	99.9	0.1	99.9	0.1	93.1	6.9	86.4	13.6
Org	16.7	83.3	7.1	92.9	2.7	97.3	0.01	99.9
Accuracy	95.09		97.87		94.32		90.31	

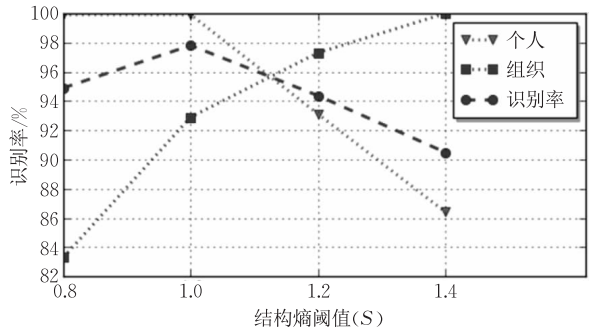


图 8　CNIM 准确率

由图 8 可知，随着识别阈值的增加，个人用户准确率总体呈下降趋势，而组织用户总体呈上升趋势. 在本文设置的结构熵识别阈值中，当识别阈值为 1.0 时，取得最大识别准确率 97.8%. 该方法识别的其他评价指标(EIs)，如表 6 所示.

表 6　CNIM 评价指标(EIs)　（单位：%）

EIs	$IT(S)$-E_u							
	0.8		1.0		1.2		1.4	
	Ind	Org	Ind	Org	Ind	Org	Ind	Org
Precision	99.9	83.3	99.9	92.9	93.1	97.30	86.4	99.9
Recall	93.6	99.7	97.2	99.7	98.8	85.18	99.9	74.9
*F*1-Score	96.7	90.8	98.5	96.2	95.9	90.84	92.7	85.7

由表 6 可以看出，随着识别阈值增加，内容规范化识别方法识别个人用户的召回率呈上升趋势且均在 90%以上，而组织用户呈下降趋势. 当识别阈值为 1.0 时，该识别方法中 *F*1-Score 取值均在 96%以上. 由此，可以说明该识别方法能有效识别社交网络中用户的组织-个人身份.

4.2.5　多媒体内容识别(MCIM)

本文识别中通过对用户(个人及组织用户约为 5160)图片内容的分析，在已有的基于 PCA 脸识别算法基础上，匹配图片人物是否为同一个人. 本文设置不同图片比率值 P_u 作为组织-个人身份的识别阈值，识别准确率如表 7 所示.

本文根据实验结果，得到不同图片比率识别阈值 $IT(P_u)$ 下的识别准确率，具体如图 9 所示.

表 7 多媒体(图片)内容实验准确率 (单位:%)

User	$IT(P_u)$-P_u							
	0.01		0.02		0.03		0.04	
	Ind	Org	Ind	Org	Ind	Org	Ind	Org
Ind	93.8	6.25	90.5	9.5	89.3	10.7	86.3	13.7
Org	49.4	50.6	66.7	33.3	83.3	16.7	91.7	8.3
Accuracy	72.18		61.90		53.01		47.32	

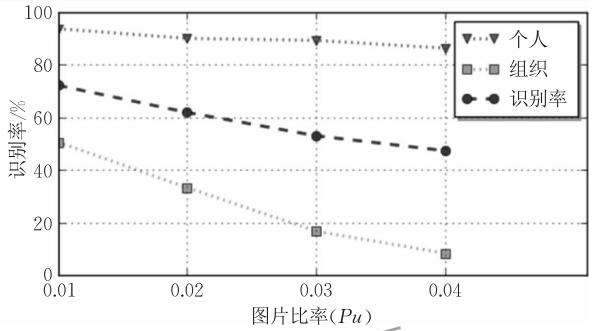


图 9 MCIM 准确率

由表 7 和图 9 可知,随着阈值的增加个人用户和组织用户识别准确率均呈下降趋势. 当识别阈值为 0.01 时,取得用户组织-个人身份识别的最大识别准确率 72.18%. 该方法识别的其他评价指标(EIs),如表 8 所示.

表 8 MCIM 评价指标(EIs) (单位:%)

EIs	$IT(P_u)$-P_u							
	0.01		0.02		0.03		0.04	
	Ind	Org	Ind	Org	Ind	Org	Ind	Org
Precision	93.8	50.6	90.5	33.3	89.3	16.7	86.3	8.3
Recall	65.5	89.0	57.6	77.8	51.7	60.9	48.5	37.8
*F*1-Score	77.1	64.5	70.4	46.6	65.5	26.2	62.1	13.6

由表 8 可知,随着识别阈值增加,多媒体内容识别方法识别个人用户和组织用户的召回率均逐渐下降. 当识别阈值为 0.01 时,该识别方法取得最佳识别结果,其中识别个人用户时 *F*1-Score 为 77.1%,识别组织用户时 *F*1-Score 为 64.5%.

4.2.6 时间序列内容识别(TSCIM)

在本文时间序列内容识别中,通过对用户时间序列的分析,得出该用户时间序列的字母序列并求出相近函数值,进而识别出该用户的组织-个人身份. 本文在不同粒度下(gr=24、48、96)进行实验,得出用户身份识别的结果. 同时,为了得出式(9)中参数 μ 的不同取值对实验结果的影响,本文计算不同 μ 值下的识别准确率. 最终得出当 μ=0.5 时取得最佳用户身份识别结果. 其中,用户识别准确率结果如表 9 所示. 不同参数 μ 下的识别准确率,如图 10 所示.

表 9 TSCIM 准确率 (单位:%)

参数_μ 粒度 gr	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
24	73.62	73.97	74.59	74.83	74.34	75.55	74.17	77.70	77.64	76.00	73.82
48	68.19	67.59	69.10	71.60	72.62	78.31	76.37	76.39	75.52	75.09	72.55
96	63.85	65.59	64.50	70.80	74.38	80.85	79.26	76.59	70.91	66.45	64.40

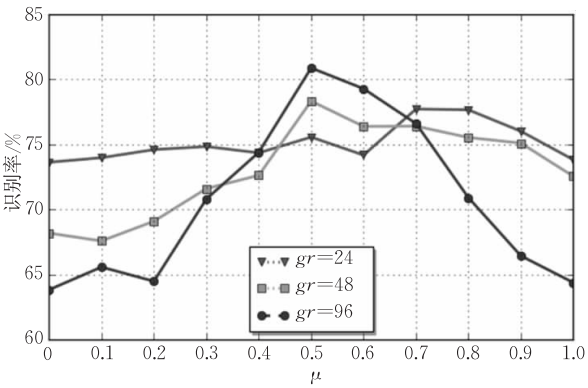


图 10 TSCIM 识别准确率率

由图 10 可知,时间序列用户身份识别在 μ=0.5 时总体取得最佳实验结果,识别准确率为 80.85%,且准确率随着粒度 gr 的增加而变大. 同时,本文计算时间序列内容识别方法在识别用户组织-个人身份中的其他评价指标. 根据识别结果,得到个人用户和组织用户实验结果评价指标,这两种身份用户评价指标比较,如图 11 所示.

由图 11 可知随着粒度 gr 的增大,召回率、*F*1-Score 值整体均呈增大的趋势.

在实验中,识别个人用户的召回率随着 μ 值的增加而上升,μ=0.5 时 *F*1-Score 取得最大值 85.77%. 同时,识别组织用户的召回率在 μ=0.5 时取得最大值 63.44%,*F*1-Score 值为 70.76%. 由此,验证了本文所提时间序列识别方法能够识别出社交网络用户的组织-个人身份.

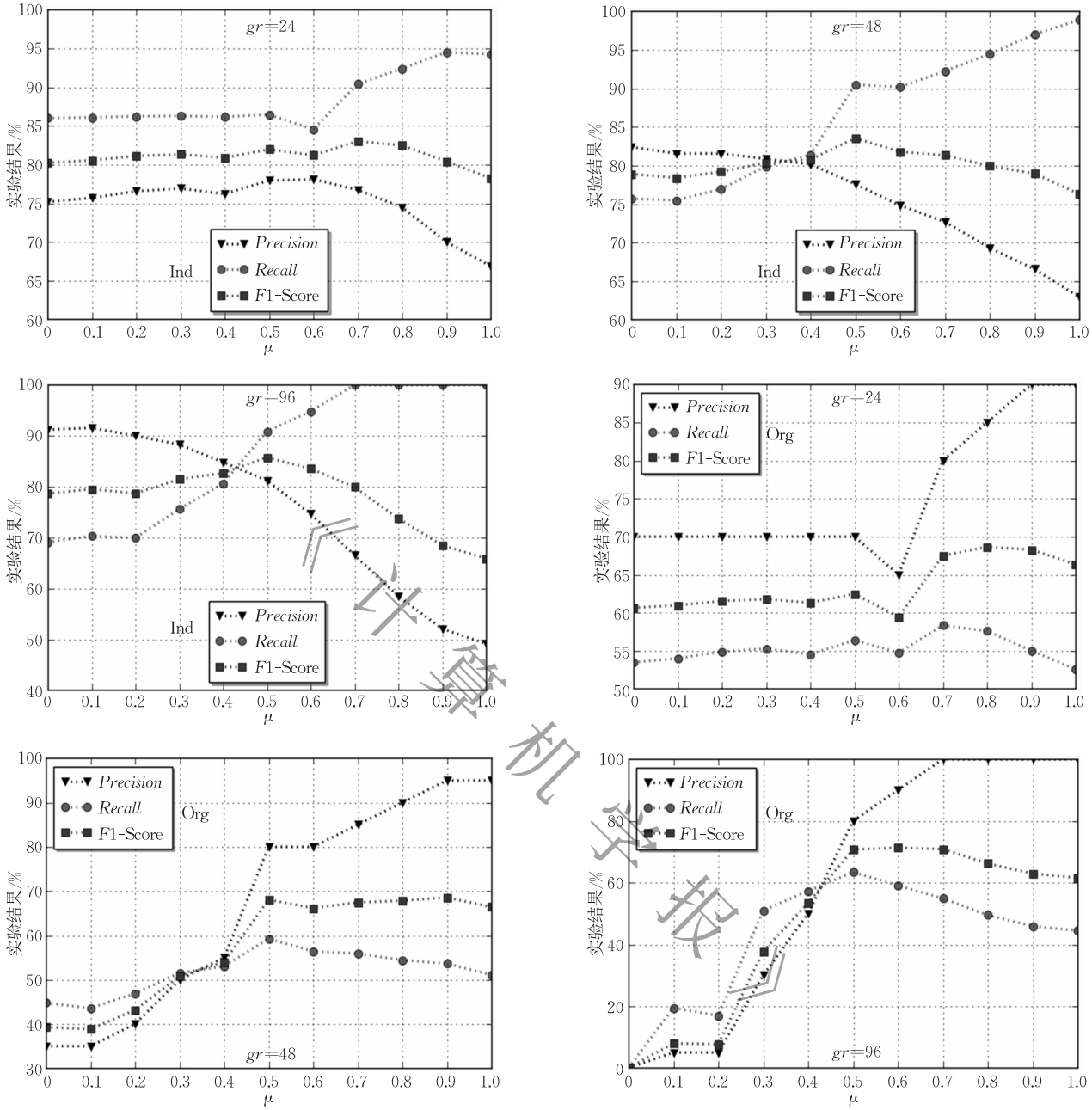


图 11　TSCIM 精确率、召回率、F1-Score 比较

4.3　实验结论

在实验中，本文通过新浪微博数据集验证了本文提出的 5 种社交网络用户组织-个人身份识别方法的有效性和可行性. 在这 5 种识别方法中，内容（主题）复杂特性识别以及内容规范化识别方法是从用户内容的宏观角度进行的分析和识别，能够刻画出用户社交内容的宏观特性. 因此，这两种方法识别效果比较明显，识别率也最高. 口语人称识别方法和多媒体内容识别方法则是从用户内容的微观角度分析和识别，能够反映出用户发表内容中的微观特性. 时间序列识别方法是对用户社交行为方面的规律分析，是对用户社交行为的一种刻画. 本文各个识别方法在准确率最高情况下的评价指标对比，如表 10 所示.

表 10　识别方法评价指标比较　（单位：%）

识别方法	EIs(%)(Ind/Org)			
	Precision	*Recall*	*F*1-Score	*Accuracy*
CPIM	Null	Null	Null	67.4/61.0
CCIM	93.3/78.8	91.5/82.7	92.4/80.7	89.10
CNIM	99.9/92.9	97.2/99.7	98.5/96.2	97.87
MCIM	93.8/50.6	65.5/89.0	77.1/64.5	72.18
TSCIM	81.21/80.0	90.88/63.44	85.77/70.75	80.85
PM	57.84/63.22	79.38/37.98	66.92/47.46	59.40

由表 10 可知，通过概率模型识别用户的组织-个人身份，准确率仅为 59.4%. 与本文提出的方法

相比，本文方法识别效果较理想.本文提出的方法中CCIM、CNIM、TSCIM方法实验结果准确率均达到80%以上.由此可以说明，本文所提出的社交网络用户组织-个人身份识方法是可行和有效的，能够识别出社交网络中用户的组织-个人身份.

5 总结与展望

在社会生活中，每个人相对其他人或事物来说，都拥有所在环境下的一种身份.在社交网络中，用户的身份同样会反映在社交网络的虚拟空间中.由于用户本质属性的不同，组织用户与个人用户在社交内容方面会呈现不同的特征.

本文提出了5种社交网络用户组织-个人身份识别的方法，从内容的微观角度、宏观角度对用户的文本内容、多媒体内容进行分析、处理，并对用户社交行为在时间上的规律进行总结分析，识别出用户的组织-个人身份.本文仅对常见的用户社交内容信息进行处理，可适用于不同的社交网络平台，也为分析用户社交内容提供了新的度量方法.从这两个角度讲，本文用户身份识别方法具有一定的广泛适用性.同时，为其他具体身份用户的识别提供了帮助和方法借鉴.此外，本文提出的用户社交内容度量方法，也可作为用户特征进行其他方面的研究和应用，如用户同一性判别、用户社区划分或聚类等.

在社交网络中，对用户身份的识别研究，可从以下三个主要方面进行：(1)社交网络内容角度，包括用户不同形式的内容信息(文本、图像、音频和视频等)以及网络用户的各种属性信息；(2)社交网络结构角度，从结构上来看，可以分为“点”(节点)、“线”(链接)、“面”(社区)的粒度，本文社交网络中用户组织-个人身份的识别还可充分利用两种身份用户的结构特点进行识别；(3)社交网络用户时间序列角度，即社交网络用户在线时间的角度.社交网络中用户的身份识别，可以充分利用不同身份用户在线时间的特点进行分析和识别.总之，在研究具体应用问题或不同社交环境时，可能涉及到以上三个方面.例如，关键用户可以分为结构关键用户、内容关键用户和时间关键用户.结构关键用户可能是由自身特点决定的，也可能是由该用户的在路径的关键位置决定的，还可能是其在社区中或社区间的重要位置决定.内容关键用户是因为它在社交网络上发布的内容影响力大.时间关键用户的典型代表是具有社交网络瘾的用户.

下一步研究可包括：(1)本文提出的社交网络用户身份识别方法，仅仅是根据用户在社交网络中产生的内容进行识别，具有一定的局限性.因此，可以增加用户身份识别的维度和选择，如网络结构、用户描述信息等；(2)本文提出的社交网络用户身份识别方法虽然能够识别出一个用户的个人-组织身份，但还有一些问题需要解决，如怎样更好地设置其中的识别阈值，以及如何更好地组合这些判别依据.此外，在社交网络用户身份识别研究中，可以从增加特征选择，引入新的特征度量，提高优化组合技术以及跨领域知识的结合应用等方面进行分析和研究.

参 考 文 献

[1] Agarwal N, Liu H. Blogosphere: Research issues, tools, and applications. ACM SIGKDD Explorations Newsletter, 2008, 10(1): 18-31

[2] Newman M E J. The structure and function of complex networks. SIAM Review, 2003, 45(2): 167-256

[3] Hu Kai-Xian, Liang Ying, Xu Hong-Bo, et al. A method for social network user identity feature recognition. Journal of Computer Research and Development, 2016, 53(11): 2630-2644(in Chinese)
(胡开先，梁英，许洪波等. 一种社会网络用户身份特征识别方法. 计算机研究与发展，2016，53(11)：2630-2644)

[4] Liu Dong, Wu Quan-Yuan, Han Wei-Hong, et al. User identification across multiple websites based on username features. Chinese Journal of Computers, 2015, 38(10): 2028-2040(in Chinese)
(刘东，吴泉源，韩伟红等. 基于用户名特征的用户身份同一性判定方法. 计算机学报，2015，38(10)：2028-2040)

[5] Wang Zhong-Hua, Han Zhen, Liu Ji-Qiang, et al. ID authentication scheme based on PTPM and certificateless public key cryptography in cloud environment. Journal of Software, 2016, 27(6): 1523-1537(in Chinese)
(王中华，韩臻，刘吉强等. 云环境下基于PTPM和无证书公钥的身份认证方案. 软件学报，2016，27(6)：1523-1537)

[6] Raad E, Chbeir R, Dipanda A. User profile matching in social networks//Proceedings of the International Conference on Network-Based Information Systems. Takayama, Japan, 2010: 297-304

[7] Cortis K, Scerri S, Rivera I, et al. An ontology-based technique for online profile resolution//Proceedings of the International Conference on Social Informatics. Kyoto, Japan, 2013: 284-298

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

[9] Wang Yun-Hong, Zhu Yong, Tan Tie-Niu. Biometrics personal identification based on iris pattern. Acta Automatica Sinica, 2002, 28(1): 1-10(in Chinese)
(王蕴红，朱勇，谭铁牛. 基于虹膜识别的身份鉴别. 自动化学报，2002，28(1)：1-10)

[10] Wang Liang, Hu Wei-Ming, Tan Tie-Niu. Gait-based human identification. Chinese Journal of Computers, 2003, 26(3): 353-360(in Chinese)
(王亮，胡卫明，谭铁牛. 基于步态的身份识别. 计算机学报，2003，26(3)：353-360)

[11] Brainard J, Juels A, Rivest R L, et al. Fourth-factor authentication: Somebody you know//Proceedings of the ACM Conference on Computer and Communications Security (CCS 2006). Alexandria, USA, 2006: 168-178

[12] Yang S H, Long B, Smola A, et al. Like like alike: Joint friendship and interest propagation in social networks//Proceedings of the International Conference on World Wide Web(WWW 2011). Hyderabad, India, 2011: 537-546

[13] Luo S, Morone F, Sarraute C, et al. Inferring personal economic status from social network location. Nature Communications, 2017, 8: 15227

[14] Subbian K, Aggarwal C C, Srivastava J. Querying and tracking influencers in social streams//Proceedings of the ACM International Conference on Web Search and Data Mining. San Francisco, USA, 2016: 493-502

[15] Naskar D, Mokaddem S, Rebollo M, et al. Sentiment analysis in social networks through topic modeling//Proceedings of the 10th Edition of the Language Resources and Evaluation Conference(LREC). Portorož, Slovenia, 2016: 46-53

[16] Sixto J, Almeida A, López-De-Ipiña D. Improving the sentiment analysis process of spanish tweets with BM25//Proceedings of the International Conference on Applications of Natural Language to Information Systems. Salford, UK, 2016: 285-291

[17] Hu Xia, Tang Lei, Tang Ji-Liang, et al. Exploiting social relations for sentiment analysis in microbloggin//Proceedings of the 6th ACM International Conference on Web Search and Data Mining. New York, USA, 2013: 537-546

[18] You Quan-Zeng, Luo Jie-Bo, Jin Hai-Lin, et al. Robust image sentiment analysis using progressively trained and domain transferred deep networks//Proceedings of the National Conference on Artificial Intelligence. Austin, USA, 2015: 381-388

[19] Chao Lin, Tao Jian-Hua, Yang Ming-Hao, et al. Long short term memory recurrent neural network based multimodal dimensional emotion recognition//Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. New York, USA, 2015: 65-72

[20] Li Jian-Xin, Wen Jiang-Feng, Tai Zhen-Ying, et al. Bursty event detection from microblog: A distributed and incremental approach. Concurrency & Computation Practice & Experience, 2016, 28(11): 3115-3130

[21] Wang Zhen-Jun, Wang Shu-Hui, Zhang Wei-Gang, et al. Social content based latent influence propagation model. Chinese Journal of Computers, 2016, 39(8): 1528-1540(in Chinese)
(王祯骏，王树徽，张维刚等. 基于社交内容的潜在影响力传播模型. 计算机学报，2016，39(8)：1528-1540)

[22] Chen Feng, Chao Wen-Han, Zhou Qing, et al. Convolution tree kernel based sentiment element recognition approach for Chinese microblog. Computer Science, 2014, 41(12): 133-137(in Chinese)
(陈锋，巢文涵，周庆等. 基于卷积树核的中文微博情感要素识别. 计算机科学，2014，41(12)：133-137)

[23] Liu Jin-Long, Wu Bin, Chen Zhen, et al. Research on influence of Micro Blogging based on field division. Computer Science, 2015, 42(5): 42-46(in Chinese)
(刘金龙，吴斌，陈震等. 基于领域划分的微博用户影响力分析. 计算机科学，2015，42(5)：42-46)

[24] Cao Peng, Li Jing-Yuan, Man Tong, et al. Detecting near duplicate messages in Twitter. Journal of Chinese Information Processing, 2011, 25(1): 20-27(in Chinese)
(曹鹏，李静远，满彤等. Twitter 中近似重复消息的判定方法研究. 中文信息学报，2011，25(1)：20-27)

[25] Chen Hai-Yan, Liu Chen-Hui, Sun Bo. Survey on similarity measurement of time series data mining. Control and Decision, 2017, 32(1): 1-11(in Chinese)
(陈海燕，刘晨晖，孙博. 时间序列数据挖掘的相似性度量综述. 控制与决策，2017，32(1)：1-11)

ZHANG Shu-Sen, Ph. D. candidate. His research interests include data mining, social computing.

LIANG Xun, Ph. D., professor, Ph. D. supervisor. His research interests include data mining, business intelligence and social computing.

MI Bao-Tong, Ph. D. candidate. His research interests include Internet of Things and social computing.

ZHAO Ji-Chao, M. S. candidate. Her research interests include internet public opinion spread and social computing.

ZHOU Xiao-Ping, Ph. D. candidate, lecturer. His research interests include data mining and social computing.

Background

In social networks, the identity of the user is an essential part of their social activities, and it may be explicit or implicit and more or less expressed in the content generated by the social networks. Thus, the identity of the user can be identified by analyzing the user-generated content. It has practical significance to identify the organization and individual identity of the user in the social networks for society, enterprise and social networks research. It connects the online virtual world with our real world, and helps us to make full use of the network. In this paper, we take the users in social networks as the research object, and put forward the problem of organization-individual identification for users in social networks. We can provide a basis for the further research and development of social networks and the identification and analysis of more specific identity of social network users.

At present, the research on user identity in social networks mainly focuses on network identity authentication, the same user judge, privacy protection and so on. There is relatively little research on user identification. In view of the text content and multimedia content user posted, we study the user's identity in social networks, which is divided into organization and individual (or personal). Through the measurement of the user's colloquial, content (theme) complexity and the normalization of text content, simultaneously considering the analysis of user's picture characteristic in multimedia content and time series content, we propose the organization and individual identity recognition (identification) methods from different angles that can be handled by computer. Thereby identifying whether the user is an organization or individual identity in social networks. To verify the feasibility and effectiveness of the proposed methods, select the micro blogging dataset from the real social network to experiment. The experimental results show that the proposed method can effectively identify the organization-personal identity of the user in the social network.

This paper is supported by the National Natural Science Foundation of China under Grant Nos. 71531012, 71271211, 71601013, the Natural Science Foundation of Beijing under Grant Nos. 4172032, 4174087, and the Science and Technology Planning Project of Beijing Education Commission under Grant Nos. SQKM201710016002.