DOI：10.7652/xjtuxb201312004

模式无关的社交网络用户识别算法

叶娜[1,2]，赵银亮[1]，边根庆[2]，李健[3]，何箐[1,2]
（1.西安交通大学电子与信息工程学院，710049，西安；2.西安建筑科技大学信息与控制工程学院，710055，西安；
3.陕西电力信通有限公司，710075，西安）

摘要：针对识别社交网络用户时存在的模式不一致问题，提出了基于分块和二部图的用户识别算法。该算法通过将传统分块算法中的属性值精确匹配扩展为无模式信息下的属性值近似匹配，避免了传统用户识别时所需的模式对齐；使用加权二部图及Kuhn Munkres（KM）最大权匹配算法进行源用户档案与待匹配用户档案间的相似度计算，解决了用户档案间属性个数不同及语义语法异构的问题。在社交网站Profilactic上采集了965个用户的公开数据，采用召回率、精确率和综合指标等评价指标对算法进行了实验评估。实验结果表明，所提算法能够不依赖模式信息进行实例级跨系统用户识别，与基于属性值精确匹配的算法相比，所提算法的召回率提高了6.2%～9.5%，综合评价指标提高了3%～4.2%。

关键词：用户识别；二部图；实例匹配；跨系统个性化
中图分类号：TP391 **文献标志码**：A **文章编号**：0253-987X(2013)12-0019-07

A Schema-Independent User Identification Algorithm in Social Networks

YE Na[1,2], ZHAO Yinliang[1], BIAN Genqing[2], LI Jian[3], HE Qing[1,2]
(1. School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China;
2. School of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China;
3. Shaanxi Electric Power Information and Telecommunication Cooperation Limited, Xi'an 710075, China)

Abstract: A user identification algorithm based on blocking and bipartite graph is proposed to solve the problem of schema inconsistency in user identification across social networks. The schema alignment is avoided by extending the precise attributes matching in traditional blocking to approximately matching without using schema information. The weighted bipartite graph and the Kuhn Munkres (KM) algorithm are employed to calculate the similarity between the source and the candidate user profiles so that the problems of different attribute numbers as well as the semantic and syntactic heterogeneity between user profiles are solved. Public data of 965 users are collected from the Profilactic website and the proposed algorithm is evaluated using the recall, precision and F-measure metrics. Experimental results show that the proposed algorithm can implement cross-system and instance-level user identification without the use of schema information, and a comparison with the method using precise attributes matching shows that the recall of the algorithm is improved by 6.2%-9.5% and the F-measure is improved by 3%-4.2%.

Keywords: user identification; bipartite graph; instance matching; cross-system personalization

收稿日期：2013-07-17。 作者简介：叶娜(1979—)，女，博士生；赵银亮(通信作者)，男，教授，博士生导师。 基金项目：国家自然科学基金资助项目(61272458)；陕西省自然科学基础研究计划资助项目(2013JM8021)；西安建筑科技大学青年基金资助项目(2013JK1189)。
网络出版时间：2013-07-19 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20130719.0910.004.html

社交网络为人们的交互提供了新的途径，人们在使用社交网络的同时也在其上留下了大量的个人信息。社交网站 Profilactic(http://www.profilactic.com)表明，一个用户通常在若干个社交网站上注册以与不同社交网站上的朋友进行交互。集成不同社交网络上的用户信息能够为个性化服务及跨领域推荐提供更全面的用户数据，也是解决冷启动问题[1]的一个重要途径。用户识别是指判断来自两个不同系统、具有不同用户模型的用户档案是否描述了现实世界中同一个用户的过程。在不同应用领域，与之等价的术语还有记录链接(record linkage)、实体识别(entity resolution)等。用户识别的准确率是决定跨系统用户信息集成效果的关键因素。

目前，在跨系统用户识别技术研究上，现有的方法主要是基于用户的属性特征进行用户相似性判定。每个注册用户都有自己的档案，如果两个用户档案相同或相似，就认为这两个用户是现实世界中的同一个人。研究者们定义了一组标识属性，识别算法将候选用户的每个标识属性的重要性因子进行组合[2]，如果某个用户组合后的重要性因子超过指定阈值，则该用户被视为匹配用户。文献[3]将用户档案表示为由若干个属性构成的向量，两个用户档案的相似性通过每个属性及其权重的乘积的线性累加和来度量，精确匹配、局部匹配和模糊匹配3种匹配函数被用于度量属性域的相似性，属性权重通过专家直觉及训练得到。尽管该算法能够达到83%的准确度，但它所考虑的属性与应用领域紧密耦合，当应用领域改变时，属性的权重必须重新进行训练。研究者们将标签和基于用户名的方法相结合以提高特定环境下用户识别算法的性能[4]。文献[5]提出了一个用户识别框架，它将来自不同社交网络的用户档案转换为 FOAF(Friend of a Friend)表示；对不同类型的用户属性采用了不同的相似性度量，属性的权重可以手动指定，也可以自动计算；如果两个用户档案的相似性得分超过指定阈值，则认为它们代表同一个人。该算法的主要问题是所使用的默认逆功能属性(Inverse Functional Property, IFP)是用户的邮件地址，而邮件地址属于用户的隐私信息，通常不能被访问。

纵观现有方法，主要是基于模式匹配的用户识别，即在进行用户识别时假设所使用的用户属性的含义以及两个社交网络上的用户属性映射关系均是已知的。然而，实际的两个社交网络其用户属性集合往往不完全相同，即使语义上一致的两个用户属性在描述上也存在语义异构及语法异构。因此，基于模式匹配的用户识别方法不得不在识别时进行两个社交网络用户档案的模式对齐或匹配[6]，这虽然有助于加快后期用户档案的相似性计算，但是两个社交网络的用户属性未必存在一一映射关系，而且需要本体或语义词典以及相应模式匹配算法的支持，从而增加了用户识别的复杂性。

本文针对社交网络用户属性集合不相同及属性语义、语法异构问题，基于分块和二部图，提出了无需模式匹配的实例级用户识别算法。该算法首先使用基于近似匹配的分块方法对待识别用户档案集进行过滤，采用扩展的 Dice 系数法作为属性值相似性度量，使用二部图描述待匹配的两个用户档案，并借助完全二部图最大权匹配 KM 算法计算两者间的相似度，最后根据设置的阈值来进行用户识别成功与否的判定。实验结果表明，该算法在召回率、准确率和综合评价指标上都得到了改善。

1 问题定义

社交网络用户识别的目标是在另一具有不同用户档案模式的社交网络用户档案集中，找到与所考虑用户档案描述同一现实世界用户的用户档案。为了避免对用户档案模式信息的依赖，我们将用户档案定义为一个属性值的集合。用户识别就是在待识别的社交网络用户档案集中找到与所考虑用户档案匹配的用户档案。

定义 1 用户档案 U 是一个属性值的集合，$U=\{v_i \mid n\geqslant i\geqslant 1\}$，其中 n 是此用户档案所包含属性值的个数。

定义 2 两个用户档案 U_1 和 U_2 匹配是指 U_1 和 U_2 描述现实世界中同一个用户。

定义 3 用户识别是指在待识别用户档案集 S 中找到与用户档案 U_s 匹配的用户档案 U_o，这里 U_s 称为源用户档案，U_o 称为目标用户档案。

由于社交网络用户档案数据量大并且模式不一致，在进行社交网络用户识别时需要解决以下问题：①如何对待识别用户档案集进行过滤，确定候选用户档案集，以减少待识别的用户档案数；②如何计算两个具有不同模式的用户档案间的相似度。

2 候选用户档案集的确定

社交网络是一个庞大的用户信息空间，在这种大规模数据集上进行用户识别时通常会根据某些属性对待识别用户档案集进行过滤，得到候选用户档

案集，以缩小搜索空间，提高搜索效率。本文通过延伸传统分块算法[7]的思想来确定候选用户档案集。传统分块算法使用模式信息，对同一属性的属性值进行精确匹配，将数据分成块，只对位于同一块内的数据进行比较[8]。然而，针对社交网络用户档案的模式异构与数据多样性，同一物理用户在不同社交网络上同一属性的属性值不一定相同，具有某些相同属性值的两个用户档案未必指向同一物理用户。采用基于传统分块算法的精确匹配会缩小搜索空间，但需要依赖模式信息，而且有可能会过滤掉真正匹配的用户档案。如何在缺乏用户档案模式信息的情况下确定候选用户档案集，是本文所要解决的主要问题之一。

为了解决该问题，本文将传统分块算法基于模式信息的属性值精确匹配扩展为无模式信息下的属性值近似匹配，该方法是在文献[8]提出的无模式信息属性值精确匹配思想的启发下所做的延伸，即对应于现实世界同一用户的两个用户档案不论属性是什么，至少有一对属性值是相同的。该方法将“相同”扩展为“相似”，使得在缺乏模式信息的情况下对待识别用户档案集进行过滤。例如，对图 1 描述的源用户档案 U_s 和待识别用户档案集 S，假设经过模式匹配后，属性“screen name”与属性“display name”语义一致，属性“locale”和属性“location”语义一致。

U_s:

screen name	full name	locale	interest
mark	Mark Smith	London	football

s:

	full name	display name	interest	birthday	location
p_1	Mark Brown	maek	novel	19830329	Plymouth, UK
p_2	Mark SW	mark2011	football computer	19820402	Regent Street, London,UK
p_3	Mark Smith	markS	football, music	19920501	Plymouth,NK,
p_4	Mark White	markWhite	football, novel	19741203	Regent Street, London,UK

c_1: $\{p_3\}$
c_1: $\{p_1, p_3\}$
c_1: $\{p_1, p_2, p_3\}$

图 1　3 种确定候选用户档案集的方法

若采用传统分块算法，根据属性“full name”进行分块，则生成的候选用户档案集为 $C_1\{p_3\}$；若不考虑模式信息，采用属性值的精确匹配，则 U_s 的第一个属性值“mark”和 p_1 的第二个属性值匹配，U_s 的第二个属性值“Mark Smith”和 p_3 的第一个属性值匹配，从而生成候选用户档案集 $C_2\{p_1, p_3\}$；当不考虑模式信息并采用属性值的近似匹配时，当 U_s 的任一属性值和某一候选用户档案的任一属性值的相似度超过设定阈值时，则将该候选用户档案加入候选用户档案集。设属性相似度阈值为 0.7，经过模糊匹配生成的候选集为 C_3，由于精确匹配得到的候选集中的用户档案必满足某两个属性值的相似度为 1，因此 C_3 中必包含 C_2 中的候选用户档案，此外，经过属性值相似度计算，U_s 的第一个属性值和 p_2 的第一个属性值的相似度为 0.75，则 p_2 也应加入候选用户档案集中，故得到 C_3 为 $\{p_1, p_2, p_3\}$。根据人为判断，在 S 中，p_2 与 p_3 是最有可能与 U_s 匹配的用户档案，但基于精确匹配的方法由于对属性值相似性要求过于严格，只将 p_3 加入候选用户档案集，而基于近似匹配的方法则适当放宽了属性值相似性条件，使得在搜索空间减小的前提下不遗漏任何可能匹配的候选用户档案。

3　用户档案的相似性度量

来自两个不同社交网络的用户档案，其属性描述及属性个数往往不同，基于模式匹配的识别方法不得不进行模式对齐。针对该问题，本文采用二部图及其最大权匹配算法进行用户档案的相似性计算。

3.1　二部图

设 $G=(V,E)$ 是一个无向图，如果顶点 V 可分割为两个互不相交的子集 (A,B)，并且图中的每条边 (i,j) 所关联的两个顶点 i 和 j 分别属于这两个不同的顶点集 $(i \text{ in } A, j \text{ in } B)$，则称图 G 为一个二部图。如果对于任意两个顶点 $i\in A$ 和 $j\in B$，(i,j) 都是 G 的一条边，则 G 为完全二部图。如果二部图的每条边上附有一个权值，则该二部图称为加权二部图。给定一个二部图 G，在 G 的一个子图 M 中，M 的边集中的任意两条边都不依附于同一个顶点，则称 M 是一个匹配。如果一个匹配中，G 中的所有顶点都在匹配边上，则称此匹配为完全匹配，也称作完备匹配。对于一个加权二部图，若存在一个匹配，其所有边上的权值之和最大，则称该匹配为此加权二部图的最大权匹配。

3.2　属性值的相似性度量

由于本文所提算法是基于无模式的用户识别，不需要考虑具体属性及其类型，因而将所有的属性值都视为字符串。字符串相似性比较的度量有很多种，包括编辑距离、Jaro 及 Jaro-Winkler 距离、Jaccard 相似度以及 Cosine 相似度等。由于 Dice 系数法[9]具有较快的运算速度，因而这里选择使用双字母组的 Dice 系数法来计算两个属性值的相似度

$$s(x,y) = 2n_t/(n_x + n_y) \quad (1)$$

式中：n_t 是两个字符串共有的双字母组的个数；n_x 是 x 中双字母组的个数；n_y 是 y 中双字母组的个数。

3.3 用户档案的相似性度量

设源用户档案 $U_s=\{u_i \mid 1\leqslant i\leqslant m\}$，候选用户档案 $U_c=\{v_j \mid 1\leqslant j\leqslant n\}$，令 U_s 中的所有属性值构成顶点集 A，U_c 中的所有属性值构成顶点集 B，从 A 中的任一顶点 u_i 到 B 中的每个顶点 v_j 都有一条边 e_{ij}，其边上的权值 w_{ij} 为 u_i 和 v_j 这两个属性值的相似度，可由 3.2 节计算求得，如此生成一个加权二部图 $G=(A+B,E,W)$，其中 E 为所有边 e_{ij} 的集合，W 为所有权重 w_{ij} 的集合。求 U_s 和 U_c 的相似度可转化为求加权二部图 G 的最大权匹配。为使 U_s 和 U_c 的相似度在[0,1]内，可将其定义为[10]

$$S(U_s,U_c) = W_{\max}/\mathrm{Max}(m,n) \quad (2)$$

式中：$W_{\max}$ 为根据二部图最大权匹配算法求得的最大权重。

求二部图最大权匹配的常用算法是 Kuhn Munkres (KM)算法，该算法能够求得完全二部图上的最大权匹配，它要求左右两边顶点数相同。对于源用户档案和候选用户档案，其属性值个数未必相同，为构造出左右顶点数相同的完全二部图，可以通过给属性值较少的用户档案添加虚拟属性值[10]，并令与其相连的边上权重为 0，图 2 给出了一个示例。

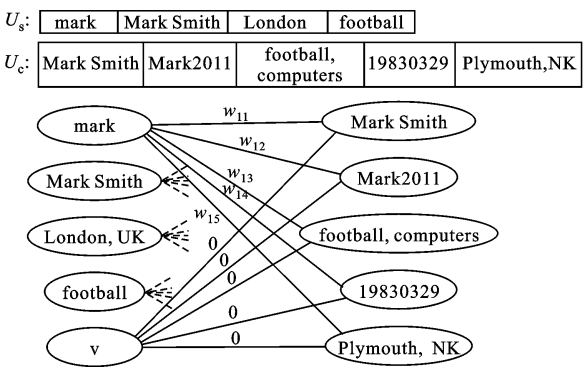

图 2 用户档案到完全加权二部图的转换示例图

具体转换过程如下。

输入 源用户档案 U_s，待识别用户档案 U_o，权值矩阵 $\boldsymbol{W}_o$(存储着 U_s 和 U_o 两两属性值间的相似度)

输出 U_s 和 U_o 对应的完全二部图

Construct(U_s, U_o, W_o)

{

$l_s=U_s$ 中属性值的个数；

$l_o=U_o$ 中属性值的个数；

$l_{\max}=\max\{l_s, l_o\}$；

创建左顶点集合 V_s 和右顶点集合 V_o；

创建 $l_{\max}\times l_{\max}$ 维的权值矩阵 $\boldsymbol{W}=\{w_{ij} \mid w_{ij}\in \mathbf{R}$ and $0\leqslant i< l_{\max}$ and $0\leqslant j< l_{\max}\}$；

以 U_s 中每个属性值创建结点并加入结点集 V_s；

以 U_o 中每个属性值创建结点并加入结点集 V_o；

将 $\boldsymbol{W}_o$ 复制到 $\boldsymbol{W}$ 的对应位置中；

if ($l_s< l_o$)

{ for ($i=l_s$; $i< l_{\max}$; $i++$)

{ 创建虚拟结点 n_i，并加入 V_s；

将 $\boldsymbol{W}$ 的第 i 行的每个元素均赋值为 0；}

}

else if ($l_s> l_o$)

{ for ($i=l_o$; $i< l_{\max}$; $i++$)

{ 创建虚拟结点 n_i，并加入 V_o；

将 $\boldsymbol{W}$ 的第 i 列的每个元素均赋值为 0；}

}

将所有从 V_s 中的顶点到 V_o 中所有顶点的边都加入到边集合 E 中；

return (V_s+V_o, E, $\boldsymbol{W}$)；

}

4 基于分块和二部图的用户识别算法

4.1 算法流程

鉴于以上分析，本文提出基于分块和二部图的社交网络用户识别算法，算法流程如下。

输入 源用户档案 U_s，待识别用户档案集 S

输出 S 中与 U_s 匹配的用户档案 U_o

(1)//确定候选用户档案集 C，构造权值矩阵 $\boldsymbol{W}$

for $i=0$ to $m-1$ //m 是 U_s 中属性值的个数

for $k=0$ to Length(S)-1

for $j=0$ to $n-1$//n 是 $U_j\in S$ 中属性值的个数

{ 计算属性值 u_i 和 v_j 的相似度 $s(u_i, v_j)$；

if $s(u_i, v_j)>\alpha$ and U_j 不在 C 中//α 为属性值相似度阈值

{ 将 U_j 放入到候选用户档案集 C 中；

计算 $\boldsymbol{W}_k$；

break；}

}

(2)//计算 U_s 与 U_j 的相似度，求得相似度最大的候选用户档案 U_o

$v_{\mathrm{similarity}}=0$；

for $j=0$ to Length(C)-1

{　//构造 U_s 和 U_j 的完全加权二部图
G=Construct(U_s, U_j, W_j)；
//使用 KM 算法计算 G 的最大权重
v_{sim}=KM(G)/Max(m, n)；
if v_{sim} > $v_{similarity}$
{$v_{similarity}$=v_{sim}；U_o=U_j；}}
(3)if $v_{similarity}$ > β//β 为用户档案相似度阈值
return U_o；

4.2　阈值选取

在根据源用户档案的每个属性值进行分块时，如果源用户档案 U_s 的属性值 u_i 和候选用户档案 U_j 的属性值 v_j 之间的相似度超过了某一阈值 α，则候选用户档案 U_j 就会被放在候选用户档案集中。当判定候选用户档案是否与源用户档案匹配时，如果两者的相似度超过了阈值 β，则算法认为两者匹配。为了确定 α 和 β 的合理取值，我们通过在社交网络上采集真实、公开的个人数据进行测试，得到了匹配率最高时的 α 和 β 的值。

4.3　时间复杂度分析

本文的用户识别算法需对待识别用户档案集 S 和候选用户档案集 C 各进行一次扫描。算法第一阶段的时间复杂度依赖于待识别用户档案集的大小 $|S|$、源用户档案的属性值个数 m 和候选用户档案的属性值个数 n，在最坏情况下第一阶段的时间复杂度为 $O(mn|S|)$。假设候选用户档案集的大小为 $|C|$，源用户档案与候选用户档案属性值个数最大值为 n，由于 KM 算法的最优时间复杂度为 $O(n^3)$，算法第二阶段的时间复杂度为 $O(|C|n^3)$，在最坏情况下候选用户档案集的大小与待识别用户档案集相同，即 $|S|=|C|$，因此该算法的时间复杂度为 $O(n^3|S|)$。由此可见本文算法的时间复杂度与待识别用户档案集的大小成线性关系，具有较好的可扩展性。

5　实验及结果分析

在实验中，测试计算机的 CPU 为 Intel COREi 53.0 GHz 4 核，内存为 3 GB，操作系统为 Window 7，采用 Java 语言进行程序开发。评价指标采用信息检索中的召回率 R、精确率 P 和综合指标 F，其计算公式为

$$R = N_c/N_a \tag{3}$$

$$P = N_c/N_s \tag{4}$$

$$F = 2RP/(R+P) \tag{5}$$

式中：N_c 是算法识别出的正确匹配的用户档案对的个数，N_a 是实验数据集中实际存在的匹配用户档案对的个数，N_s 是算法识别出的匹配用户档案数（包括正确匹配的和不正确匹配的）。

通过实验，准备验证：①阈值 α、β 与 R、P 及 F 之间的关系；②本文算法在不同类型数据集上的 R、P 及 F；③算法的执行效率。

5.1　实验数据

为了得到真实数据，我们在社交网站 Profilactic 上采集了 965 个用户的公开个人数据。Profilactic 网站上给出了它的每个用户所注册的社交网站，我们所采集的用户都在 Flickr 上进行了注册，通过 Profilactic 用户信息页上的超链接，可以转向该用户在 Flickr 上的用户信息页。据此，我们可以验证找到的等价用户是否是正确的。

5.2　实验结果及分析

为了验证第一个问题，令阈值 α 从 0.5 递增到 1，β 从 0.1 递增到 0.9。实验结果如图 3 所示。

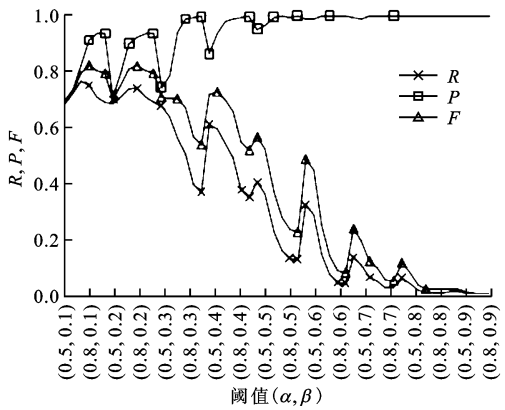

图 3　召回率 R、精确率 P、综合评价指标 F 随阈值的变化

通过实验确定，当属性值相似度阈值 $\alpha=0.8$，用户档案相似度阈值 $\beta=0.1$ 时，F 达到最大值 82.5%，R 为 75.2%，P 为 91.3%。如图 3 所示，当用户档案相似度阈值 β 保持不变，属性值相似度阈值 α 递增时，R 减小，这是由于阈值增加时，过滤后的候选用户档案集缩小（当 $\alpha=1$ 时，即为精确匹配），说明基于精确匹配的分块方法过于严格，可能会过滤掉实际匹配的用户档案。图 4 为采用精确匹配时，R、P 及 F 随阈值 β 的变化情况。由图可见，基于精确匹配的算法的 P 在不同 β 取值时普遍比较高，但 R 却普遍低于基于近似匹配的算法的值，这反映出基于精确匹配的算法在确定候选用户档案集时，可能过滤了实际匹配的用户档案。

对于第二个问题，我们从 965 个用户档案中筛

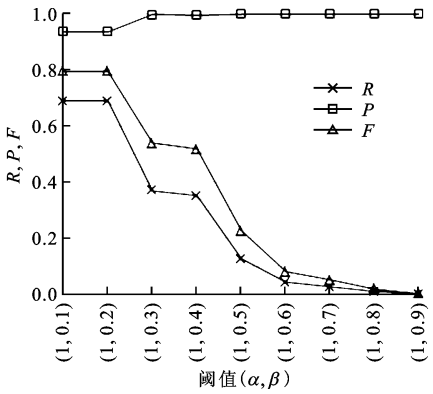

图 4 基于精确匹配的召回率 R、精确率 P、综合评价指标 F

选出其待识别用户档案集中不包括空属性值的 530 个用户档案，图 5 的实验结果表明，当源用户档案与待匹配用户档案的属性值个数相同时，其 R、P 及 F 高于属性值个数不同时的值，即属性值信息越完整算法识别的效果越优。

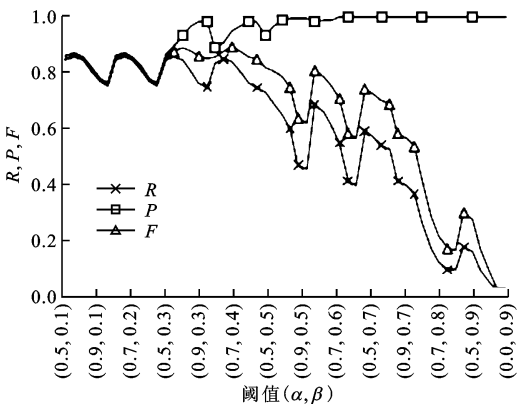

图 5 完整数据集上召回率 R、精确率 P、综合评价指标 F 随阈值的变化

此外，将本文算法与文献[11]及基于精确匹配的算法进行比较，结果如表 1 所示，与文献[11]基于模式匹配的算法相比，本文算法的 R、P 和 F 得到了显著提高。对于同样类型的数据集，基于精确匹配算法的 P 均高于基于近似匹配的算法的值，但 R 和 F 均低于基于近似匹配的算法的值，这也反映出社交网络上用户数据的不准确性与不一致性。不论是基于精确匹配的算法还是基于近似匹配的算法，属性值越全面算法的效果越优。

表 1 与其他算法的效果对比 %

算法	R	P	F
文献[11]算法	73.0	88.8	80.1
精确匹配算法(原始数据集)	69.0	93.8	79.5
本文算法(原始数据集)	75.2	91.3	82.5
精确匹配算法(完整数据集)	74.7	98.5	85.0
本文算法(完整数据集)	84.2	94.9	89.2

为了验证算法的执行效率，我们在不同规模的数据集上进行了测试，实验结果如图 6 所示，算法的执行时间与待识别用户档案集的规模成线性关系。为了测试算法执行时间与阈值间的关系，令 α 从 0.1 递增到 0.9，由于阈值 β 对算法执行时间没有影响，因此令 $\beta=0.1$。图 7 的实验结果表明，属性相似度阈值 α 越大，得到的候选用户档案集规模就越小，算法的执行时间也越少；另外，当 $\alpha=0.8$ 时，F 达到最大值，尽管这时的运行时间不是最优，但仅次于 $\alpha=0.9$ 时的最优执行时间，这也表明，当(α,β)取值为(0.8,0.1)时，算法的执行效率相对较优。

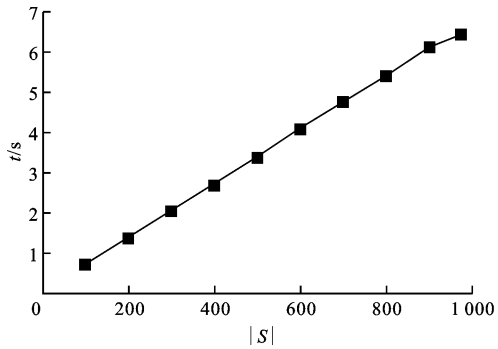

图 6 不同规模数据集上的执行时间

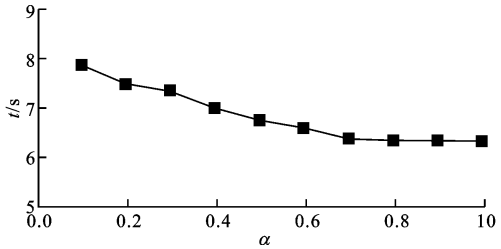

图 7 执行时间随阈值 α 的变化

6 结 论

针对社交网络用户识别时存在的模式不一致及数据异构问题，提出了在可用模式信息有限的情况下，基于分块和二部图的实例级社交网络用户识别算法。在分块思想的基础上，将基于模式信息的属性值精确匹配扩展为无模式信息的属性值近似匹配，根据用户档案的属性值相似度将待识别用户档案集进行分块，确定候选用户档案集。采用扩展的

Dice 系数法作为属性值相似性度量，为源用户档案和候选用户档案建立二部图，使用求完全二部图最大权匹配的 KM 算法计算两个用户档案的相似度。通过使用实际社交网站上的公开用户数据，得到属性值相似性阈值及用户档案相似性阈值。当候选用户档案集的最大相似度大于用户档案相似性阈值时，判定用户识别成功，具有该最大相似度的候选用户档案被判定为与源用户档案匹配。下一步工作是研究在缺少共同属性时的用户识别方法。

参考文献：

[1] BODHIT A, AMIN K. Possible solutions of new user or item cold-start problem [J]. International Journal of Mathematics and Computer Research, 2013, 1(3): 123-128.

[2] CARMAGNOLA F, CENA F. User identification for cross-system personalisation [J]. Information Sciences, 2009, 179(1): 16-32.

[3] VOSECKY J, HONG D, SHEN V Y. User identification across multiple social networks [C]// Proceedings of the First International Conference on Networked Digital Technologies. Piscataway, NJ, USA: IEEE, 2009: 360-365.

[4] IOFCIU T, FANKHOUSER P, ABEL F, et al. Identifying users across social tagging systems [C]// Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. Palo Alto, California, USA: AAAI, 2011: 1-4.

[5] RAAD E, CHBEIR R, DIPANDA A. User profile matching in social networks [C]// Proceedings of the 13th International Conference on Network-Based Information Systems. Piscataway, NJ, USA: IEEE, 2010: 297-304.

[6] MARTINEZ -VILLASENOR M L G, GONZALEZ-MENDOZA M. Process of concept alignment for interoperability between heterogeneous sources [C]// Proceedings of the 11th Mexican International Conference on Advances in Artificial Intelligence. Berlin, Germany: Springer-Verlag, 2013: 311-320.

[7] BAXTER R, CHRISTEN P, CHURCHES T. A comparison of fast blocking methods for record linkage [C]// Proceedings of the First Workshop on Data Cleaning, Record Linkage and Object Consolidation. New York, USA: ACM, 2003: 25-27.

[8] PAPADAKIS G, IOANNOU E, NIEDERÉE C, et al. Efficient entity resolution for large heterogeneous information spaces [C]// Proceedings of the 4th ACM International Conference on Web Search and Data Mining. New York, USA: ACM, 2011: 535-544.

[9] WHITE S. How to strike a match [EB/OL]. (2010-05-03) [2012-05-25]. http://www.catalysoft.com/articles/StrikeAMatch.html.

[10] 李默涵，王宏志，李建中，等. 一种基于二分图最优匹配的重复记录检测算法 [J]. 计算机研究与发展，2009，46(S2)：339-345.
LI Mohan, WANG Hongzhi, LI Jianzhong, et al. Duplicate record detection method based on optimal bipartite graph matching [J]. Journal of Computer Research and Development, 2009, 46(S2): 339-345.

[11] CARMAGNOLA F, OSBORNE F, TORRE I. User data distributed on the social web: how to identify users on different social systems and collecting data about them [C]// Proceedings of the First International Workshop on Information Heterogeneity and Fusion in Recommender Systems. New York, USA: ACM, 2010: 9-15.

[本刊相关文献链接]

王龙翔，张兴军，朱国峰，等. 重复数据删除中的无向图遍历分组预测方法. 2013，47(10)：51-56. [doi：10.7652/xjtuxb201310009]

李清华，康海燕，苑晓姣，等. 个性化搜索中用户兴趣模型匿名化研究. 2013，47(4)：131-136. [doi：10.7652/xjtuxb201304022]

张赛，徐恪，李海涛. 微博类社交网络中信息传播的测量与分析. 2013，47(2)：124-130. [doi：10.7652/xjtuxb201302021]

陈国强，王宇平. 采用离散粒子群算法的复杂网络重叠社团检测. 2013，47(1)：107-113. [doi：10.7652/xjtuxb201301021]

薛咏，冯博琴，刘卫涛. 扩展主题图本体融合策略与算法. 2011，45(10)：13-18. [doi：10.7652/xjtuxb201110003]

豆增发，高琳. 应用粒子群优化-条件随机域的文本生物实体识别. 2010，44(12)：38-42. [doi：10.7652/xjtuxb201012008]

鲁慧民，冯博琴，李旭. 面向多源知识融合的扩展主题图相似性算法. 2010，44(2)：20-24. [doi：10.7652/xjtuxb201002005]

（编辑　武红江）