基于校园一卡通数据好友发现及应用

鲁鸣鸣，张丹，王建新
中南大学信息科学与工程学院，湖南 长沙 410083

摘要
教育大数据是近年来的研究热点。利用分布式文件系统对校园一卡通数据进行存储、预处理和分布式计算。在此基础上，提出学生生活轨迹中的相遇模型，从而挖掘学生线下社交关系。为了区分熟悉的陌生人和真正的好友，从单个学生和整个班级两个角度分析学生线下相遇行为，既可以挖掘好友关系（包括比较孤立的学生），又可以为校园班级社群管理提供数据支撑。实验结果表明，挖掘出的社交关系比较符合实际情况。

关键词
社群发现；相遇模型；线下社交关系

中图分类号：TP181　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2017021

Smart-card based campus friend mining and its applications

LU Mingming, ZHANG Dan, WANG Jianxin
School of Information Science and Engineering, Central South University, Changsha 410083, China

Abstract
Recently, educational big data has become a hot topic. A distributed file system to store, preprocess, and analyze campus card data was adopted. Based on it, a student encounter model has been proposed, so as to mine students' offline social relations. To distinguish real friends from familiar strangers, the offline social relations for either individual students or classes were analyzed. Through these two perspectives, the students' offline encounters was analyzed, which can not only extract social relationship between friends (including the isolated students), but also provide data support for the campus class management. The experimental results show that the mined social relations reflect the real relationship.

Key words
community discovery, encounter model, offline social relation

1 引言

“数据驱动学校，分析变革教育”的大数据时代已经来临，利用教育大数据技术为教育教学决策提供有效的支持将成为未来教育的发展趋势。世界知名高校、机构和企业纷纷启动了教育大数据相关研究计划，从研究内容来看，国外教育大数据研究工作主要围绕学生信息系统和在线学习行为，综合运用统计学、机器学习和数据挖掘的技术，研究如何促进学习者学习的有效性，并根据学生能力和实际需求，提供个性化教育方案[1]。我国高校对在校生实行统一集中的住校食宿管理，这一特点使得我国的教育大数据不仅需要关注学习的有效性和个性化教育，还需要在心理健康和学生校园管理方面提供更好的指导。集中式的管理方式给我国的教育大数据发展带来了新的机会，提供了国外教育机构难以获取的学生线下行为数据。例如，不少高校都有校园一卡通系统，其数据能反映学生在食堂就餐、去图书馆、打热水、洗浴等生活轨迹信息。基于一卡通数据进行分析，国内已有不少研究成果[2-4]。

通过分析智慧校园数据来反映大学生学习生活状态，是帮助学校掌握学生动态一个有效途径。但是，很多研究工作是基于学生上网数据分析学生线上行为，线上数据不能准确反映学生的线下行为。虽然已有不少对学生线下行为分析的工作，但是绝大部分都只对线下行为数据进行统计分析，没有对学生线下社交关系进行挖掘。

笔者根据学生刷卡数据建立学生相遇模型，通过关联规则挖掘学生社交关系，并从学生角度、班集体角度两个方面反映学生线下社交情况。

笔者基于中南大学校园一卡通数据集，通过用户就餐刷卡数据记录，构建相遇模型，从而发现用户刷卡相遇，利用关联规则理论，计算相关度量（如支持度、可信度等），找出经常一起刷卡的学生，进而生成线下社交关系网络图、挖掘好友、发现社群。为验证相遇模型分析发现好友的准确度，对56名学生进行了访谈和问卷，发现好友准确度超过90%。

本文的主要贡献如下：

- 基于校园一卡通线下数据，定义相遇模型，挖掘学生线下社交关系；
- 将模型推广到一般场景，适用于一切具体相遇场景数据用户的好友挖掘；
- 挖掘出班级社交关系，找出“孤立”学生以及小团体活动的学生，及早发现潜在的孤独症学生，引导更健康的班级社群关系，为学校更好管理学生提供数据支撑。

2 相关工作

随着在线社交网络的蓬勃发展和线上用户的急剧增长，以交友、信息共享等为目的的社交网络迅速成长为人们传播信息、表述观点、发泄情绪的理想平台。很多研究者基于用户线上数据分析线上用户社交关系[5,6]，但是线上社交分析没有办法很好地反映用户线下行为。

相关研究者对大学生心理健康的研究及调查表明[7-9]，当代我国大学生心理健康形势比较严峻，相当一部分学生心理上存在各种不良反应和适应障碍，有的甚至到了较严重的程度，心理健康问题已经影响到大学生的健康成长与发展。因此，积极运用一切手段开展大学生心理问题的预防工作，是全社会的一项重要任务。但是大学生心理问题预防是一项长期的工程，需要对每个学生跟踪阶段性的状态，因此，挖掘学生线下社交，了解学生交友动态，引导

和帮助学生积极健康学习、交友和生活是非常有意义的。

当前研究者对校园一卡通数据做了多方面的研究。罗华群[10]提出结合各单位的申请和一卡通系统的消费数据来审核助学金和贷款发放工作，比单纯由各院系收集学生们写的申请会更加合理和高效，也会更客观；陈建兵[11]利用一卡通系统数据辅助辨别贫困生；金培莉等人[12]提出从学校后勤、人事等部门以就餐人数、月均消费情况和洗浴情况3组数据分析来支撑学校决策等。研究人员基于校园一卡通数据及各类校园数据进行多方面分析挖掘，但是到目前为止，还没有研究者基于校园一卡通数据这类线下活动数据来分析学生社交关系。

3 整体思路

笔者的目标是通过校园一卡通数据挖掘学生好友以及学生社交状态，而好友总是倾向于一起活动，如一起去食堂就餐、去图书馆等，而这些线下行为在一卡通刷卡记录里得以反映，因此，可以通过分析一卡通刷卡记录的时空相关性发现学生之间一起活动的行为。同时，按照一起刷卡这个行为挖掘好友时，存在“熟悉的陌生人”这种情况，即有相似作息行为习惯却并非好友的陌生人，因此提出相遇模型解决这一问题。

利用校园一卡通刷卡数据进行相遇建模，计算学生刷卡相遇情况，然后以学生刷卡相遇数据计算相关支持度和可信度，从而发现关联规则，进一步分析学生之间的关联性，最后找出每个学生关联的学生，并将学生线下社交关系以社交网络图的形式呈现。为了更准确地计算相遇，对部分地点的刷卡机（POS机）聚类，使相遇更加精准。

学生好友人群分析查找工作涉及数据预处理、关联分析、关系挖掘等多个过程[13]，这其中也有很多关键难点需要克服和解决，主要集中在以下几个方面。

- 数据处理分析过程涉及多个阈值，很容易影响分析结果。在POS机聚类阶段有两个阈值，针对学生每一次就餐存在多次刷卡的情况，对每次就餐刷卡的POS机序列建立任意两个POS机间的一次计数，并且采用一个月的数据进行聚类，因此两个阈值取值区间比较大，对应值的选取比较容易且对结果影响较小。再者，在学生相遇模型建立阶段，给定时间阈值dis，将在dis时间内刷卡的学生视为相遇，dis是根据学生刷卡平均占用时间来确定的，使得相遇统计结果更理想。
- 由于人与人之间的关系随着时间的推移也会改变，仅通过刷卡数据很容易漏掉熟悉的人。按照学生关联规则分析，根据实际刷卡规律提出3个限制条件，若满足3个条件中的一个，便被视为熟悉人群，这样就有效地解决了这类问题。
- 相似性行为的陌生人由于生活规律极其相似，容易被划分到熟悉人群。因为有相似的就餐习惯，很容易被错分，笔者尽量采用多种方式度量考察，避免这类陌生人被划分到熟悉人群。在预处理过程中对POS机聚类时，把食堂POS机按照楼层分开，这样能很好地排除部分相似就餐规律的陌生人。
- 部分人群极少使用校园卡，此类学生无法通过一卡通数据发现线下社交。后期笔者会加入更多类型的数据（如上网数据、宿舍刷卡等），进一步分析学生社交。

4 具体实现

4.1 数据初步预处理

数据集共有304个表，包括用户基本信

息、食堂信息、刷卡记录数据（包括卡号、刷卡时间、刷卡食堂地点、POS机编号等）等，共有用户173 490人，刷卡记录时间是2015年1月1日—12月24日，消费刷卡数据约5 014万条。数据集包含大量的字段，需从原始数据中筛选出相关字段，同时直接过滤掉异常记录。由于数据量较大，在单机上处理困难，笔者搭建了分布式Hadoop集群，将数据放在集群上进行存储和处理。将数据存放在Hadoop分布式文件系统（Hadoop distributed file system，HDFS）上，方便后续处理和分析。**表1**为筛选之后的字段，其中，*account*表示用户唯一识别字段，*des*表示刷卡机所在地点编码，*poscode*表示刷卡机编码，*time*表示刷卡时间。

表1 初步预处理后数据样例

account	*des*	*poscode*	*time*
84496	254	22	20150918 12:05:57
84496	254	87	20150918 12:10:33
84497	271	10	20150915 12:25:55
84497	271	35	20150915 12:22:28
84497	271	88	20150915 12:21:4
…	…	…	…

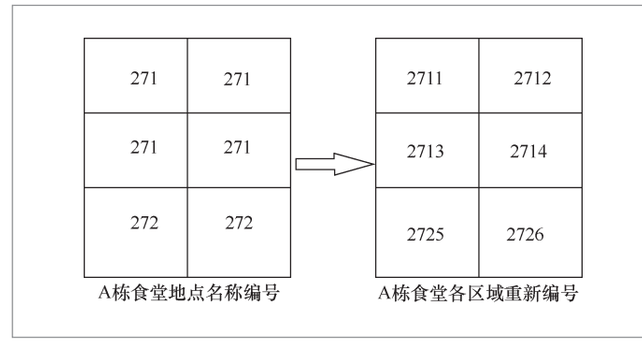


图1 A栋食堂各区域聚类重新编号

4.2 刷卡机聚类

4.2.1 模式计数

食堂刷卡是学生最典型的刷卡行为，数据集中消费刷卡数据记录为5 014万条，其中食堂刷卡记录为4 378万条，超过90%的刷卡记录为食堂刷卡。由于中南大学的食堂通常都是整栋楼，每栋楼里有若干个食堂，而刷卡地点*des*与食堂并非是一一对应的，数据中*des*往往对应若干个属于同一栋楼的食堂。若不做处理，容易把陌生人当成好友。因此，通过对所有食堂的所有POS机聚类，将每栋食堂每一区域的POS机分开。如**图1**所示的A栋食堂，其中箭头的左侧为A栋食堂的结构，第一层左右两个食堂的*des*编号均为272，而第二和第三层均为271。在实际的场景中，学生每次就餐在同一栋楼的多个不同食堂都刷卡的概率比较低，因此，可将每个学生每次就餐所刷的POS机归为同一区域的POS机。这样精确区分刷卡地点能更准确地估算相遇事件。其他刷卡记录与预处理后的食堂刷卡数据一样，均使用相同的方法计算相遇行为事件，因此，本文均以学生食堂刷卡数据为例进行说明。

为了高效率地处理数据，笔者使用并行数据处理高级查询语言Pig[14]对一卡通数据进行处理。

算法1：生成POS关系对

```
input=DataBag  //一个用户的所有刷卡记录
tuples=LinkedList( );
recordF = null  //一次就餐的第一次刷卡
white (DataBag.hasNext( ))
  recordN = DataBag.next( )  //一次刷卡记录
    if (list == null)
      recordF = recordN
      list.add(recordN.pos)}
```

```
else
    if (recordN.time−recordF.time<L
      &&recordN.des == recordF.des)
      for:eachpos in list
        tuples.add(des+eachpos+ “,”
        +des+recordN.pos)
        if(!list.contains(recordN.pos))
          list.add(recordN.pos)
    else
      list = null
      recordF = recordN
  return tuples
```

选取一个月的一卡通学生就餐数据，用Pig的Group方法将数据以用户账号分组，并将每个用户的刷卡交易数据以时间排序，排序后的数据作为输入传给算法1。输入的数据为*input*，*list*表示学生一次就餐刷过的POS机列表，*L*表示一个时间阈值，如果两次刷卡为同一次就餐，那么时间间隔必须在*L*内，最后返回的*tuples*表示用户每次就餐刷过的POS机集合内任意两个POS机记录，即很多个<*despos*1,*despos*2>对，然后用Pig自带的COUNT方法统计*despos*对出现的次数*times*，数据格式为：{*despos*1,*despos*2,*times*}。

表 2 *despos* 对计数数据样例

*despos*1	*despos*2	*times*
25410	25456	1 563
25427	25456	10 114
27211	27212	24 005
27211	27213	25 621
27246	27255	7 585
27246	27281	15
27226	27245	4 152
…	…	…

4.2.2 无向图聚类

通过同一用户一次就餐刷卡的模式计数统计本次刷卡的POS机列表，从而使该列表的POS机两两计数（又称*despos*对计数）值为1次，将多个用户多次就餐刷卡产生的相同*despos*对的计数值求和（见**表2**），*despos*1和*despos*2代表两个不同POS机的编号，*times*表示两POS机对应的计数和，计数值越大，表明这两个POS机属于*des*食堂同区域的可能性越大。给定一个阈值时间*T*，当*despos*对计数*times*>*T*时，则可视为POS机属于同一层。

对这些*despos*对构造无向图，每个*despos*对被视为一个节点，*times*>*T*，表示*despos*对对应的节点存在边。

图2是由*despos*对构造的无向图，其中*T*=200，当*times*>*T*，*despos*对的两个节点构成边，从而实现POS机聚类成多个簇，无向图存在*k*个连通分量，每个连通分量都是一个簇，对每个簇重新编号（按照1,2,3,…,*k*编号），再将每个簇内节点按照*des*和簇编号组合成新编号*posid*，每个簇内编号都相同，这样就把同一餐饮楼的每个区域食堂POS机分开了。

4.3 数据二次预处理

根据对POS机的聚类，每个食堂的任意POS机*despos*都有一个对应的*posid*，并且属于食堂同一区域的POS机的*posid*相同，将初步预处理的刷卡交易记录数据（见**表1**）的*des*和*poscode*用新对应的*posid*代替。

学生每次就餐刷卡次数是一次或者多次，设置时间阈值，若每个用户连续两次刷卡时间间隔在这个阈值之内，就视为同一次就餐刷卡。用Pig进行预处理，同样按照用户分组，将每个用户数据用自定义的方法处理，该方法将属于同一次就餐刷卡

的数据记录合并成一条数据记录，最后将记录返回给Pig，并保存在HDFS上。**表3**为预处理后部分数据样例，*account*t表示学生卡号，*num*表示一次就餐刷卡次数，*stime*、*endtime*分别表示一次就餐第一次刷卡时间、最后一次刷卡时间。

4.4 学生关联规则

关联规则是形如*X*->*Y*的蕴含式[15,16]，通过*X*可以推出*Y*，类比本文的学生对象以刷卡相遇为依据的关联规则，即由学生*X*刷卡可以推导学生*Y*也很有可能在同一地点刷卡。关联规则的强度可以用其支持度*S*和置信度*C*度量，在这里，定义*X*的刷卡次数为*X*的支持度，用*S*（*X*）表示，支持度*S*（*X*->*Y*）表示学生*X*和学生*Y*在同一地点同一时间区间内刷卡的次数。置信度又称可信度，*X*->*Y*的可信度为：

$$C(X->Y)=\frac{S(X->Y)}{S(X)} \qquad (1)$$

常见的关联规则挖掘算法（如Apriori算法、FP-Tree算法等[17-19]）对频繁项集进行挖掘，事先对规则剪枝，80%以上的规则会被丢弃。这些算法更多地关心出现频率较高的对象，高频对象才是关联挖掘的主体。

由于学生用户数量较多，使用上述算法对学生关联分析挖掘，提取规则的数目达到指数级，整体的计算规模极大，并且很多学生用户无法被挖掘。另外，学生关联分析挖掘的每个学生对象都作为一个主体，它更多地是在查找二项集，每位学生在一定程度上是同等重要的。因此，本文采用相遇模型来计算可信度，从而挖掘学生关联。

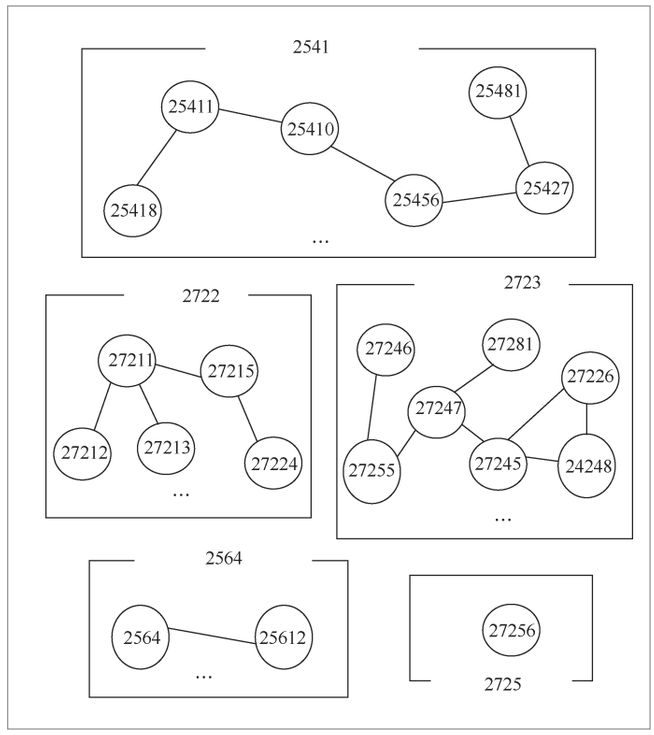


图 2 刷卡地点重新编号

表 3 二次预处理后数据样例

account	*posid*	*date*	*stime*	*endtime*	*num*
100618	27212	20150901	17:37:00	17:40:01	7
100618	25830	20150902	07:51:35	07:51:35	1
100618	27212	20150902	11:31:37	11:33:05	4
100618	27212	20150902	11:40:55	11:40:55	1
100618	27212	20150902	17:27:21	17:29:45	4
...	...	...	...	...	...

4.5 关联分析过程及相遇模型

4.5.1 学生刷卡数据

为了确定相遇的时间窗口阈值，对学生每次就餐平均的刷卡时间进行了统计，如**图3**所示。从**图3**可看出，大部分学生在1~2 min可以打完饭菜，超过150 s才打完饭菜的学生很少，只有1%。因此，在150 s的基础上适当放宽一点，设定相遇时间窗口为180 s。

4.5.2 相遇模型

为了更好地进行学生关联分析，建立了学生相遇模型，即规定在同一相遇时间窗口内、在同一食堂刷卡的记录为学生在该食堂刷卡相遇的记录。

为了统计学生的相遇记录，以1个月为时间周期，学生对象(X, Y)，i表示第i个周期，分以下3步进行处理。

步骤1 统计每个周期内每个学生的刷卡次数$S_i(X)$，其中一次就餐多次刷卡视为一次刷卡，*account*为学生X的卡号，数据格式为{*account*, $S_i(X)$}，数据文件命名为*i*_count.csv。

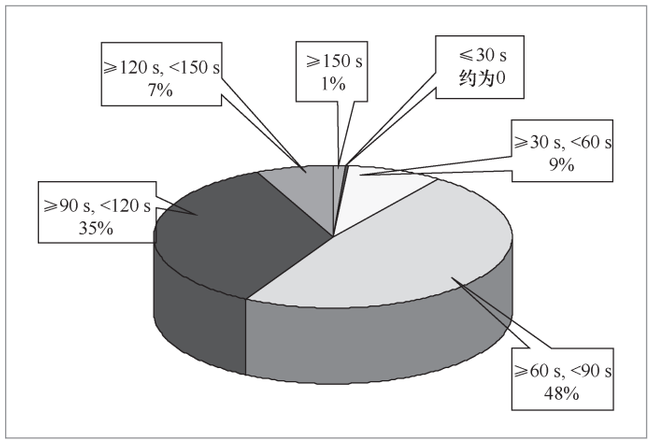


图 3 学生每次就餐平均刷卡时间

步骤2 用Pig对预处理后的数据按照(*posid*, *date*)进行Group分组，将每组数据传给算法2。该算法根据相遇定义，将所有相遇的学生形成两两记录，其中max$P$表示某用户前后关联人数，*dis*表示时间阈值，设定max$P$和*dis*两个参数同时满足条件才符合相遇。由第4.5.1节分析得知：每个学生平均刷卡时间在3 min之内的人数超过99%，从而确定了时间阈值*dis*。再统计两两记录在这个周期内出现的次数，数据格式为{*account*1, *account*2, *meettimes*, *month*}，数据文件命名为*i*_meettimes.csv。

步骤3 使用Pig将*i*_meettimes.csv文件和*i*_count.csv文件进行第一次join left连接，得到每组{*account*1, *account*2, *meettimes*}中*account*1在*i*个月中刷卡次数$S_i(X)$；再继续和*i*_count.csv进行一次join left连接，得到对应*account*2在*i*月中刷卡次数$S_i(Y)$。那么可信度$C(X->Y)=meettimes/S_i(X)$，$C(Y->X)=meettimes/S_i(Y)$。

算法2：生成相遇记录

```
input=DataBag
tuples=LinkedList();
line[maxP] =new Person() //line指相遇人群队列
indexS,indexE = 0 //line的开始索引和结束索引
values = input.iterator()
flag = false //标记是否为第一条数据
while ( values.hasNext())
  person = values.next() //用户一条刷卡数据
    if (indexS == indexE&&flag)
      indexS =(indexS+1)% maxP
    line[indexE].account = person.account
    line[indexE].stime = person.stime
    while (indexS ! = indexE) //修改开
```

始索引

```
        if(line[indexE].stime−line[indexS].stime<dis)
            break;
        indexS = (indexS+1)%maxP
    for(i=indexS;i!=indexE;)
        tmp1=line[i].account
        tmp2=line[indexE].account
        if ( tmp1>tmp2 )
            tmp1 = line[indexE].account
            tmp2 = line[i].account
        tuples.add(tmp1,tmp2)
        i = (i+1)%maxP
    indexE = (indexE+1)%maxP
    flag = true
return tuples
```

4.5.3 关联规则

根据相遇定义和算法2，进一步统计学生刷卡相遇情况，并计算出相关的支持度和可信度。最后，利用支持度和可信度两个度量标准衡量学生对象间的关联性。

设置最小支持度阈值*minsup*，对于单个学生对象*X*或者*Y*，其相关支持度必须大于*minsup*，因为支持度很低的规则可能只是偶尔出现。当$S(X)$或者$S(Y)$很低时，说明学生*X*或者*Y*刷卡的次数很低，这也意味着关联性在这种情况下难以反映出来。此外，若$S(X->Y)$很低，说明*X*、*Y*的相遇极具偶然性。

可信度的度量通过规则进行推理具有可靠性，对于给定规则$X->Y$，可信度越高，*X*与*Y*一起刷卡的概率越高，所以可信度在一定程度上反映彼此相遇的强度，从而间接反映两者关系的关联性强度。设置最小可信度阈值*minconf*，其相关可信度必须大于*minconf*，因为如果$C(X->Y)$或者$C(Y->X)$太小，说明他们联系强度太弱。

在满足$S(X)>minsup$和$S(Y)>minsup$的前提下，*X*与*Y*满足下面至少一种情况即可被划分为关系密切，其中，*i*表示第*i*个周期。

- $C_i(X->Y)>minconf\&\&C_i(Y->X)>C_i(X->Y)/2$。
- $\sum_{i=a}^{b} S_i(X->Y)>S_{min}$。
- $(\sum_{j=1}^{n} Meet_j(XY))>K$。

由于关联规则作出的推理并不必然蕴含因果关系，且因果关系通常更多地涉及长期出现的联系，所以在上述第2个条件中，*a*,*b*表示周期，指在第*a*~*b*个周期期间，*X*、*Y*相遇的总次数大于一个设定的值S_{min}时，可以认为*X*、*Y*为关系密切。第3个条件中，$Meet_i(XY)$的取值为0或者1，当第*j*个周期$S_j(X->Y)>minsup$，即*X*、*Y*在该周期相遇的次数大于*minsup*次时，$Meet_i(XY)$取值为1，否则取值为0，那么经过*n*个周期，对$Meet_j(XY)$求和，值大于*K*次，就说明这两者关系密切。对于第1个条件，当$C_i(X->Y)>minconf$时，*X*与*Y*两个人若为相识，则彼此之间的可信度$C_i(X->Y)$和$C_i(Y->X)$应该相差不大，所以在分析*X*与*Y*是否为好友时，还需要满足条件$C_i(Y->X)>C_i(X->Y)/2$。

上述涉及两个阈值*minsup*、*minconf*的确定。其中最小支持度阈值*minsup*是为了过滤因刷卡行为较少、无法通过刷卡相遇来发现好友关系的用户，*minsup*的值取决于一个周期的时间长短，本文以1个月为一个周期，令$minsup=((1/n)\sum S(k))\times 10\%$，*n*表示用户量，$S(k)$表示用户*k*在一个周期内的刷卡次数。对于最小可信度阈值*minconf*的确定，根据经验给定*minconf*的初始值，利用采集的数据作为训练样本（信息包括：卡号、好友列表、所属班级等），对比实验结果计算罚函数，罚函数值越小，*minconf*值越优，通过不断调整*minconf*值，使罚函数值不断减少至可接受范围，最终确定当前最佳*minconf*值。

5 实验结果

中南大学校园一卡通数据包含了学生所有刷卡行为记录，将前面部分预处理技术及相遇模型应用到中南大学一卡通数据集，构建相遇模型，然后利用关联规则理论计算相关度量，最后发现学生好友关系。

为了验证之前提出的学生好友关系验证方法，通过调查问卷和访谈的形式获取真实的线下好友关系。问卷涉及内容包括用户卡号、性别、在校好友列表、所在宿舍等，并选择在5个宿舍楼（3栋男寝、2栋女寝）门口进行了为期5天的随机问卷调查，去掉信息不完整的数据后最终保留56名用户信息，其中包括37名男生，19名女生。此外，还通过联系院系辅导员，争取了3个班级的配合，分别对2个班级（共86名学生）进行了访谈，完成此次班级社群关系挖掘。

本文第4.5.2节已经完成了支持度和可信度的计算，并且第4.5.3节定义了两个学生对象关系密切需要满足的条件，由此将计算出的单个学生对象支持度以及两两学生对象的支持度和可信度，按照是否满足关系密切的条件，判定任意两个学生对象是否关系密切。

为更好地反映学生线下社交关系，本文从两个角度挖掘学生社交关系，分别是以团体为分析对象的社交网络和以个体为中心的社交网络，再通过可视化的方面呈现，从而更准确地评估分析结果与线下社交真实度。

5.1 单个对象的社交网络

以被查询的学生对象为中心，挖掘其线下社交关系，从而发现该学生线下社交情况。即对于每个被选择的学生对象X，由关联规则分析理论找出与X满足关系密切条件的所有学生对象集合A。

图4中account1和account2为两个用户4个周期与其他学生相遇情况的热力图，其中横坐标卡号表示不同用户，纵坐标表示周期，利用相遇模型计算的支持度和可信度来展现，其中方格越大，相遇概率越高，颜色越深，相识的概率越大。从**图4**可以很明显看出，4个周期中，account1用户好友关系比较稳定，而account2相对于account1用户好友关系没那么固定。

用户的活跃程度可以间接地根据用户的朋友圈规模以及用户的朋友的朋友圈规模来衡量。**图5**为3个学生的线下刷卡相遇

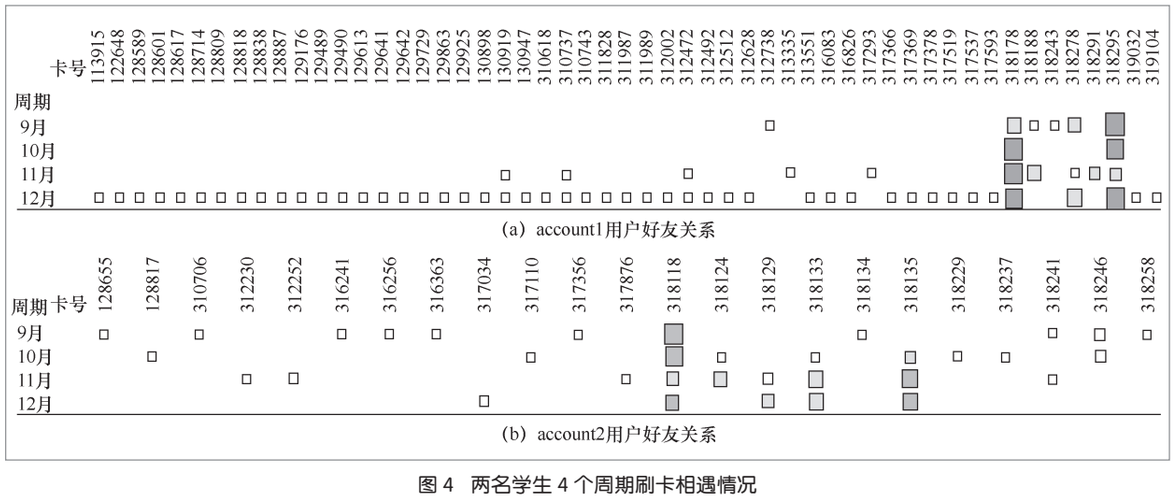
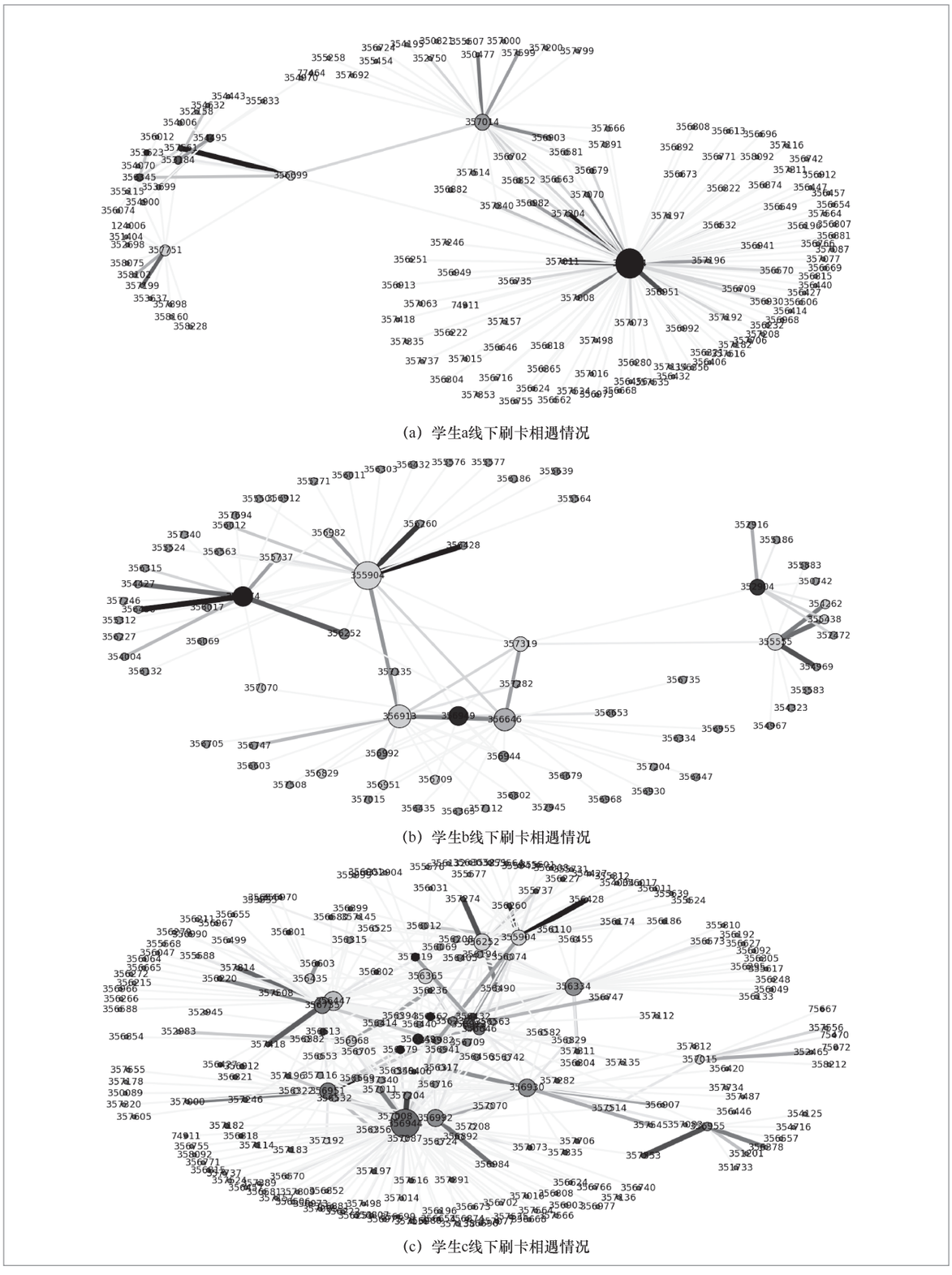


图 4 两名学生 4 个周期刷卡相遇情况



(a) 学生a线下刷卡相遇情况

(b) 学生b线下刷卡相遇情况

(c) 学生c线下刷卡相遇情况

图 5 3 个不同用户各自线下相遇情况

情况，每个节点表示一个用户个体，节点大小表示用户的可能好友的个数，节点间的边表示两个个体关系，边的粗细程度指两个用户的相遇程度，边越粗越可能关系密切。其中3个社群表示3个用户的好友以及各自好友的好友的两层传递的刷卡相遇情况，社群大小反映了该用户活跃程度，很显然，学生c比较活跃。

5.2 班级社群关系网络

一个班集体社交关系研究作为一类典型的社区发现，具有十分重要的意义。班集体内部的社交关系，反映了每个学生班内的交友情况，间接地为学校引导学生走向更健康的学习生活方式提供了数据支撑。

图6是两个不同班级线下社交关系网络图，**图6**分别为中南大学计算机专业1401班和信息安全专业1401班同学之间社交情况，其中最大节点代表非本班的学生，其他每个节点表示一名学生，不同班级呈现出社交网络差别很大。计算机专业1401班明显班内学生联系密切，主要集中在本班，信息安全专业1401班的班内联系不紧密，班内还有小群体以及单独人群。

5.3 实验评估

笔者利用问卷收集的56名用户信息以及3个班级的访谈问卷信息，从学生对象和班级对象进行分析，并根据分析的结果对比真实调研情况。**图7**为56名用户的准确率

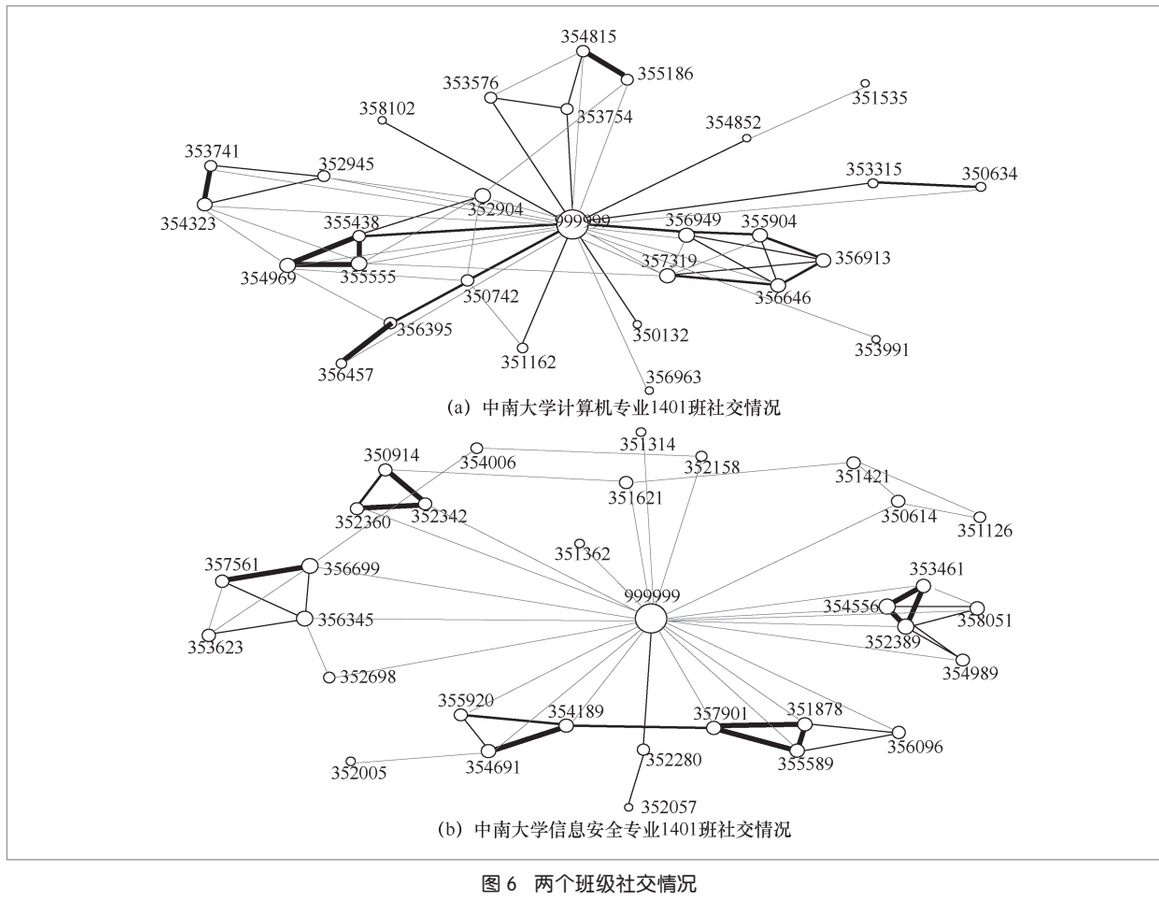


图 6 两个班级社交情况

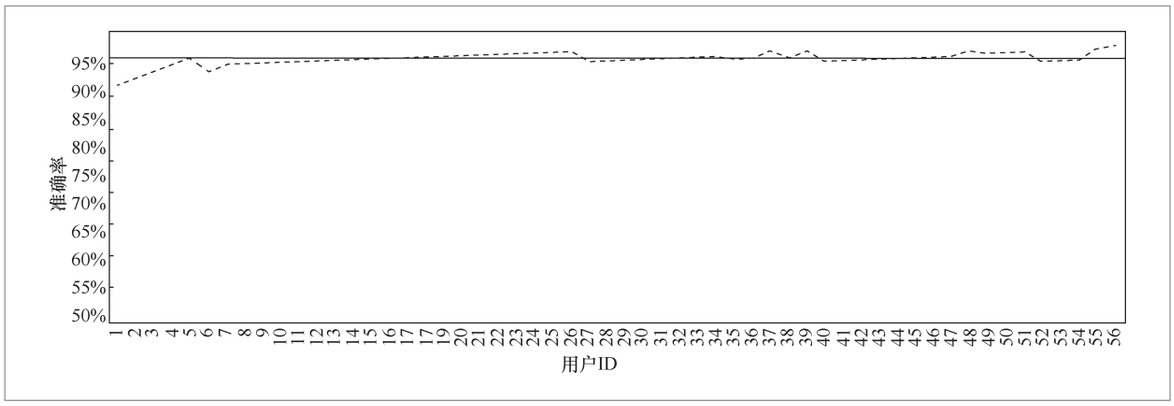


图 7 好友识别准确率

的情况，纵坐标表示用户准确率，横坐标代表不同用户，此处准确率=实验好友正确的个数/实际好友个数，虚线代表正确率曲线，实线为90%准确率参考线。从实验结果可看出，笔者提出的好友挖掘方法对好友的识别率有90%左右。

将3个班级的社交网络图分别放在对应班级访谈，并将实验结果的正确程度划分为完全不符合、部分不符合、基本符合和完全符合4个级别，如图8所示，横坐标表示不同班级，纵坐标表示各班级分析结果的不同符合程度所占比例，由图8可知，班级代号为1、3的两个班基本符合和完全符合所占比例之和超过90%，代号为2的班级同样接近90%，由此可见班级社交关系挖掘比较准确。

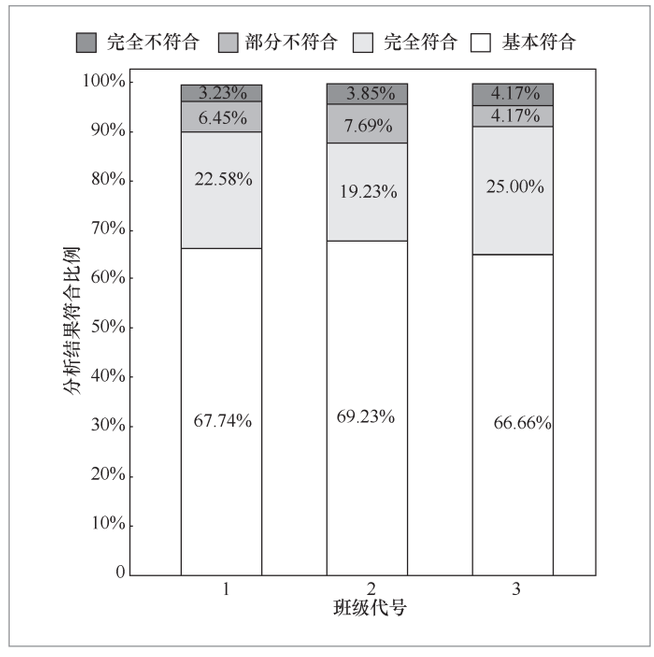


图 8 3 个班级符合程度

6 结束语

笔者建立相遇模型来统计学生相遇情况，并通过关联规则理论挖掘学生好友关系，从而找到任意学生及与其关系密切的人群。通过这种方式挖掘学生线下社交关系，有诸多方面应用。首先，通过这种方式找到学生关系密切人群，从而排除这类学生患有孤独症的可能。其次，以班级为单位进行社群发现，了解班级间学生现实生活交友交流状态，发现不利学生个人发展的现象，如固定小团体交友现象等，继而将这种班级社群关系向更健康的社群关系引导，并且为校园线下社交关系挖掘研究添加新内容。

笔者提出关于校园学生刷卡相遇模型，该模型不仅仅适用于校园刷卡挖掘学生线下社交，同时适用于其他场景的基于

线下活动数据发现用户线下社交。

参考文献:

[1] 徐鹏, 王以宁, 刘艳华, 等. 大数据视角分析学习变革: 美国《通过教育数据挖掘和学习分析促进教与学》报告解读及启示[J]. 远程教育杂志, 2013(6): 11-17.
XU P, WANG Y N, LIU Y H, et al. The learning innovation from the perspective of big data: an analysis of the U.S. report of enhancing teaching and learning through educational data mining and learning analytics and its enlightenment[J].Distance Education Journal, 2013(6): 11-17.

[2] 王德才. 数据挖掘在校园卡消费行为分析中的研究与应用[D]. 哈尔滨: 哈尔滨工程大学, 2010.
WANG D C. Research and application of data mining in campus card consumption behavior analysis[D]. Harbin: Harbin Engineering University, 2010.

[3] YAN Z. The research of data mining technology in campus card consumption system[C]// 2011 International Conference on Electronics and Optoelectronics (ICEOE), September 8, 2011, Dalian, China. New Jersey: IEEE Press, 2011: 265-267.

[4] 吕红胤, 连德富, 聂敏, 等. 大数据引领教育未来: 从成绩预测谈起[J]. 大数据, 2015045.
LV H Y, LIAN D F, NIE M, et al. Big data drives a new epoch of education: a case study of academic performance predication[J]. Big Data Research, 2015045.

[5] AGRAWAL R, IMIELIŃSKI T, SWAMI A. Mining association rules between sets of items in large databases[J]. ACM SIGMOD Record, 1993, 22(2): 207-216.

[6] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules[C]//20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago, Chile. New York: ACM Press, 1994: 487-499.

[7] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48-61.
SUN J G, LIU J, ZHAO L Y. Clustering algorithms research[J]. Journal of Software, 2008, 19(1): 48-61.

[8] ANDERBERG M R. Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks[M]. Manhattan: Academic Press, 2014.

[9] AIN A K, DUBES R C. Algorithms for clustering data[M]. New Jersey: Prentice-Hall, Inc., 1988.

[10] 杨媛媛. 社交网络中的关系构建和亲密度分析[J]. 石家庄学院学报, 2009, 11(6): 91-96.
YANG Y Y. An analysis of relationship formation and intimacy in social network[J]. Journal of Shijiazhuang University, 2009, 11(6): 91-96.

[11] 吴昊, 刘东苏. 社交网络中的好友推荐方法研究[J]. 现代图书情报技术, 2015, 31(1): 59-65.
WU H, LIU D S. Research on the method of friends' recommendation in social network[J]. New Technology of Library and Information Service, 2015, 31(1): 59-65.

[12] 张淑娟. 大学生心理健康状况与应对措施[J]. 中医药管理杂志, 2016(3): 30-31.
ZHANG S J. The mental health of college students and the countermeasures[J]. Journal of Traditional Chinese Medicine Management, 2016(3): 30-31.

[13] WAN L C. A research on the correlation of loneliness and social anxiety in college students[J]. Advances in Psychology, 2016, 6(4): 391-397.

[14] 李旭, 郑涵予, 卢勤. 大学生自杀意念及其影响因素分析[J]. 中国公共卫生, 2016, 32(3): 359-362.
LI X, ZHENG H Y, LU Q. Suicidal ideation and associated factors among university students[J]. Chinese Journal of Public Health, 2016, 32(3): 359-362.

[15] 罗华群. 校园一卡通数据的挖掘与应用[D]. 上海: 华东师范大学, 2009.
LUO H Q. Excavation and application of

campus card data[D]. Shanghai: East China Normal University, 2009.
[16] 陈建兵. 利用校园一卡通数据优化高校贫困生认定系统[D]. 成都: 电子科技大学, 2012.
CHEN J B. Using the campus card data to optimize the system of determining poor students in colleges and universities[D]. Chengdu: University of Electronic Science and Technology of China, 2012.
[17] 金培莉,王晓震. 校园一卡通系统决策支持实例分析[J]. 华东师范大学学报 (自然科学版), 2015(1): 525-529.
JIN P L, WANG X X. Instance analysis of campus card system decision support[J]. Journal of East China Normal University (Natural Science), 2015(1): 525-529.
[18] KANTARDZIC M. Data-mining concepts[J]. Data Mining: Concepts, Models, Methods, and Algorithms, Second Edition, 2011(1): 1-25.
[19] GATESA. Pig编程思想[M]. 曹坤, 译. 北京: 人民邮电出版社, 2013.
GATESA. Thinking in Pig [M]. Translated by CAO K. Beijing: Posts & Telecom Press, 2013.

作者简介

鲁鸣鸣(1978-),男,博士,中南大学信息科学与工程学院副教授,中国计算机学会会员,主要研究方向为深度学习及其在大数据、城市计算、移动计算和物联网方面的应用。

张丹(1991-),女,中南大学信息与工程学院硕士生,主要研究方向为大数据、数据分析。

王建新(1969-),博士,中南大学信息科学与工程学院教授,中国计算机学会高级会员,主要研究方向为计算机算法、网络优化理论、生物信息学。

收稿日期: 2016-10-25
基金项目: 国家自然科学基金资助项目(No.61232001, No.61173169, No.60903222);湖南省战略新兴产业重大科技研究计划基金资助项目(No.208GK4054)
Foundation Items: The National Natural Science Foundation of China(No.61232001, No.61173169, No.60903222), Hunan Strategic Emerging Industries Major Scientific and Technological Research Project(No.208GK4054)