社会语义网社区发现标签传递算法研究

蔡国永 林 航 文益民
(桂林电子科技大学计算机科学与工程学院 桂林 541004)

摘 要 针对在线社会网络的特性和现有社区发现算法的不足,提出一种基于语义网技术的在线社会网络社区发现算法ISLPA(Improved Semantic Label Propagation Algorithm),即一种适用于大规模在线社会网络的社区发现和标识算法。ISLPA算法对语义标签算法SemTagP进行改进,在社区划分过程中将在线社会网络视为有向加权图,通过语义网和社会化标签技术,充分结合在线社会网络丰富的语义信息和网络拓扑特征进行社区划分。ISLPA算法不需要预先设定社区数量和大小,就能实现社区发现,并能根据标签自动识别划分的社区。算法接近线性时间复杂度,具有较高的效率。通过实验表明,ISLPA算法能有效划分和标识真实在线社会网络。

关键词 社区发现,标签传播,语义网,社会化标签
中图法分类号 TP391 **文献标识码** A

Study on Label Propagation Based Community Detection Algorithm for Social Semantic Network

CAI Guo-yong LIN Hang WEN Yi-min
(School of Computer Science & Engineering, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract According to the characteristics of online social network and the shortcomings of the existing community detection algorithms, this paper proposed an improved community detection algorithm based on semantic technology——ISLPA (Improved Semantic Label Propagation Algorithm). ISLPA is suitable for discovering and identifying community structure in the large-scale online social network. It is an improved SemTagP algorithm, combining with semantic and social tagging technology. This algorithm takes advantage of the semantic information and topology features of online social network to community structure discovering. ISLPA doesn't require a priori information such as the number and size of communities while it's used to discovery community structures in large-scale online network, and it can also automatically identify the detected communities according to the tagging labels. This algorithm is much efficiency because it takes nearly linear time complexity. The experiment shows that SLAP algorithm can effectively discover and identify community structure for real online social networks.

Keywords Community detection, Label Propagation, Semantic Web, Social tagging

1 引言

在复杂网络中,大部分现实网络是不均匀网络,它们是由多个子网络组成的。这些子网内部个体间的关系紧密,而子网间个体的关系相对稀疏。这一现象在社会网络中尤为明显。Newman[1]把满足这一特征的子网结构称为社区结构(Network Community Structure)。社区结构是复杂网络最重要的拓扑特性之一。在线社会网络中,发现社区结构有助于理解用户分布和用户行为。

传统的社区发现算法可分为基于图划分的算法[2]、分裂算法[3]和凝聚算法[4,5]。然而这些算法的实现都需要预知网络规模、社区数量或社区规模等特征信息。但在实际分析中,在社区划分完成前社区数量和大小是很难确定的。因此,Raghavan提出了一种不依赖于先验信息的算法——标签传播算法(Label Propagation Algorithm, LPA算法)[6]。LPA算法时间复杂度为O(km),能高效地划分大规模网络,但是在密度低的网络结构中,LPA算法不能得到理想的社区划分结果。针对此问题,Ereteo提出基于语义标签传播的社区发现算法SemTagP[7]。

SemTagP算法采用用户使用的标签代替LPA算法中的随机分配值来初始化节点,并在标签传播过程中考虑标签间的语义相似关系。算法在模块度值不再增大时停止,共享同一语义标签的节点划分为同一社区,且通过语义标签识别社区的性质。然而该算法仍存在一些不足,这些缺陷影响了算法的执行效率和实用性。本文提出了改进的语义标签传播算法ISLPA(Improved Semantic Label Propagation Algorithm)。

本文第2节首先介绍SemTagP算法;第3节详细阐述ISLPA算法的主要思想、算法描述和性能分析;第4节通过

到稿日期:2012-04-21 返修日期:2012-08-20 本文受国家自然科学基金(61063039),广西自然科学基金(2011GXNSFA018156),广西可信软件重点实验室项目(kx201202)资助。

蔡国永(1971—),男,博士,教授,主要研究方向为软件形式化方法、社会网络、语义网技术,E-mail:ccgycai@guet.edu.cn;林 航(1988—),男,硕士生,主要研究方向为语义网技术、在线社会网络;文益民(1969—),男,博士,教授,主要研究方向为数据挖掘、社会网络。

实验分析验证 ISLPA 算法的可行性和可靠性。

2 语义标签传播算法 SemTagP

SemTagP 算法是对 LPA 算法的扩展。LPA 算法的基本思想[6]是：在初始化时，为网络中的每个节点随机分配唯一的标签；算法进行过程中，循环更新节点的标签，节点选择邻居节点中使用频率最高的标签来更新自己的标签；算法的终止条件是每个节点的标签不再变化；算法结束后，拥有相同标签的节点构成一个社区，标签的种类数表示社区的数量。与 LPA 算法中传播随机分配的标签不同，SemTagP 算法中为用户分配的标签是用户自己使用的标签，且在传播过程中考虑标签的泛化关系。通过这种传播方式，能将拥有具体标签的成员划分到同一上位词标签构成的社区中。语义标签传播过程如图 1 所示，图中 sport 为 football、golf、swim、rugby 的上位词；programming 为 Java、Python、Lisp 的上位词；且右图中同一底色的节点在同一社区内。

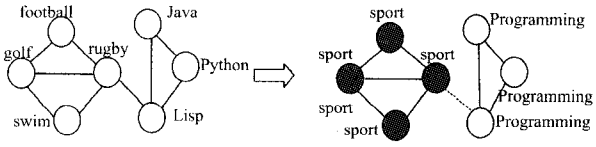

图 1 语义标签传播示意图[7]

SemTagP 算法利用 RDF 图表示网络的用户信息和标签信息，并借鉴 Leitch 和 Newman 提出的有向图社区模块度定义了 RDF 图社区模块度 $Q(G,p)$，将其用于评估 SLPA 社区划分的质量，并通过模块度的比较来决定算法是否终止。考虑到网络中数据类型的多样性和参与者之间的不同的语义关联，SemTagP 算法中将网络分析类型作为参数。SemTagP 算法的主要步骤描述如下：

(1)初始化 RDF 图中参与者节点的标签集，输入在线社会网络的 RDF 图模型 G 和待分析的关系类型 rel_Type；

(2)记录当前图模型 $old\_graph=G$；

(3)使用哈希表数据结构存储图中节点，并设置更新序列 Y；

(4)根据设置好的序列，依次使用标签更新函数 mostUsedNeighborTags($user$,rel_Type)，更新 RDF 图中的用户节点所连接的标签；

(5)比较标签更新前后网络模块度的大小，如果 $Q(G,p)<Q(old\_graph,p)$，则停止算法，输出 old_graph，否则返回步骤 2，继续执行。

其中，函数 mostUsedNeighborTags ($user$,rel_Type)用于获取 $user$ 在 rel_Type 关系下的邻居节点中使用次数最高的标签值。迭代更新每个参与者的标签时，需考虑到标签间的泛化关系。算法中标签每出现一次，该标签的上位词标签的出现次数也需要增加一次。

SemTagP 算法存在几个不足：1)算法将社会网络当作有向图处理，但真实的在线社会网络中，用有向加权图能更准确表示用户间的交互关系。2)算法仅考虑标签间的语义关系，没有充分利用用户交互关系的语义信息，例如通过用户交互关系的类型或亲密程度来评价用户间的影响力。3)共享同一标签的用户可能构成多个社区，如图 2 所示，SemTagP 算法没有对这种情况进行处理，而是简单地将同一标签的用户划分为一个社区。若将图 2 中的网络划分为两个社区，则网络模块度仅为 0.35，而将其划分为 3 个社区时，网络模块度为 0.582。4) 算法中把比较模块度作为算法的终止条件，但算法不能保证每一次迭代时模块度单调增加，因此算法会在标签传播未稳定时提前终止，如果要得到稳定的社区划分结果，则需要多次运行算法。针对这些不足，本文对 SemTagP 算法进行了改进，提出 ISLPA 算法。

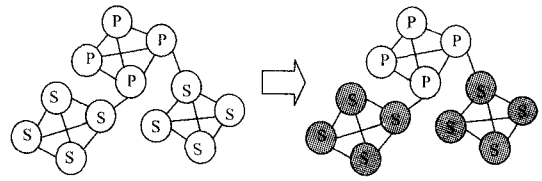

图 2 一个标签覆盖两个社区的情况

3 语义标签传播算法 ISLPA

ISLPA 算法仍使用语义标签作为传播对象，与 SemTagP 算法不同，ISLPA 算法不是在传播过程中检测标签间的语义关系，而是根据标签的语义关系和网络社区划分的尺度要求，在标签传播前预先对标签进行统一的语义处理。在语义标签传播更新过程中，算法将在线社会网络抽象为有向加权图，图中节点表示用户，边表示用户交互关系，边的权值表示某种交互关系对应的影响力。在不同交互关系下的邻居节点具有不同的影响力，因此要根据用户交互关系为边赋予权值。根据边的连接关系和边的权值大小为每个节点生成一个候选标签列表，标签列表含有语义标签及其对应的分数。各个节点选择对应列表中分数最大的标签更新原有标签。算法迭代运行，直到节点的标签不再变化时终止算法。算法中标签和用户关系的语义处理过程、标签的传播过程如图 4 所示。其中，标签的语义关系如图 3 所示，Rel1、Rel2、Rel3 表示双向关系，其对应的权值大小分别为 1、2、3；图中用无向边表示双向关系，标签的分数计算将在下一节详细介绍。

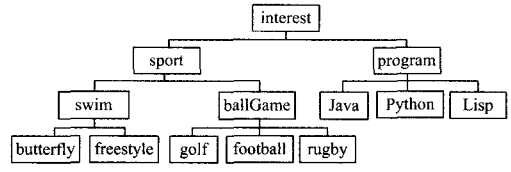

图 3 标签的层次关系

图 4 中使用两种语义处理方式对节点的初始标签进行处理。第一种语义处理方式中，使用层次较高的标签“sport”和“program”对原始标签进行替换。高层次标签其覆盖范围广，在标签传播过程中，容易形成规模较大的社区结构。在这种情况下，算法将图 3 中的示例网络分为两个社区，其中标签“sport”标识的社区规模较大。第二种语义处理方式中，算法使用了标签“program”和标签“sport”的子标签“ballGame”、“swim”对原始标签进行语义替换。使用层次较低的标签能对社区进行更细致的划分。利用第二种语义处理方式，算法将图 3 中的示例网络分为 3 个社区，第一种情况下标签“sport”标识的社区被分解为两个社区结构。

针对图 2 所示的情况，在语义标签传播结束后，算法将分析共享同一标签的节点集合的成分构成。成分是指网络中的最大关联子图。如果共享同一标签的节点只有一个成分构成，那么这些节点就构成一个社区结构；如果共享同一个标签的节点由多个成分构成，则每个成分代表一个社区结构。

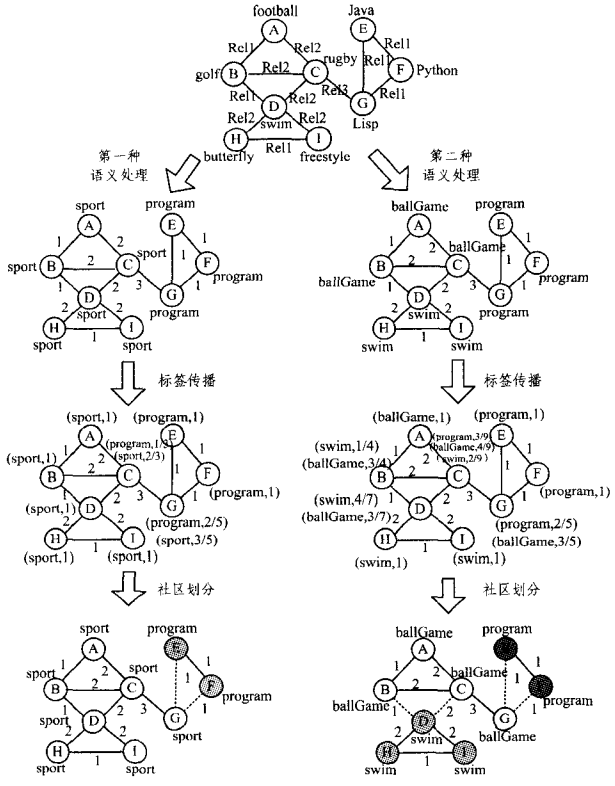

图 4　ISLPA 算法示意图

此外，算法结束后使用有向加权网络模块度 Q_{dw}[8] 来评估社区划分的质量。有向加权网络模块度可以表示为：

$$Q_{dw}=\frac{1}{W}\sum_{ij}(W_{ij}-\frac{s_i^{out}s_j^{in}}{W})\delta(c_ic_j)$$

式中，W 表示网络中所有边的权值，W_{ij} 表示边 l_{ij} 的权值，s_i^{out} 和 s_j^{in} 分别表示节点 i 和节点 j 的权值，在有向加权网络中节点的权值等于与该节点相连的指定方向边的权值之和，函数 δ 表示边的社区参与度，若节点 i 和 j 在同一社区，则 $\delta(c_ic_j)=1$；反之 $\delta(c_ic_j)=0$。

3.1　算法描述

ISLPA 算法的主要步骤描述如下：

(1)对待分析的在线社会网络 RDF 图 G 进行语义预处理，语义处理包括两个方面：一是标签语义预处理，根据标签语义上的层次关系和分析的尺度要求，将节点的初始标签替换为相应的上位词标签；二是关系类型的权值映射，根据关系的语义关系，将关系映射到不同的类型，并根据不同类型的影响力大小赋予相应的权值。将处理后的 RDF 图 G 和待分析的关系类型集合 $Type$ 作为算法的输入参数，并设置迭代计数器 $t=0$；

(2)设置 $t=t+1$；随机设置图中节点的更新序列 X；按照更新序列 X，并根据语义标签更新函数 updateTag($user$, $Type$)，依次更新用户节点。

(3)当每个节点的标签都为邻居节点中出现频率最高的标签且标签值不再变化时，停止迭代；否则返回步骤(2)，继续执行。

(4)检测共享同一标签的节点集合的成分组成。输出社区划分结果。

步骤 1 是 RDF 图的语义预处理阶段。处理中需考虑标签间的 skos:narrower 关系，“A skos:narrower B”表示 A 标签比 B 标签的含义广，称 A 为 B 的上义词，B 为 A 的下义词。含义广的标签在传播中容易形成用户数量大的社区结构。按照标签间的 skos:narrower 关系，对标签进行分层。根据社区划分的尺度要求，选中不同层次的标签，并通过检测标签间 skos:narrower，将 RDF 图中用户的初始标签替换为不同层次的上义词。此外，根据不同交互关系的影响力大小为 RDF 图中节点间不同的交互关系赋权值，生成关系类型集合 $Type$，$Type$ 含有交互关系名及其对应的权值。

步骤 2 是标签传播更新阶段。节点根据标签更新函数 updateTag($user$, $Type$)更新标签。节点根据邻居节点中相应标签的使用次数和相邻节点间关系的权值来计算邻居节点所使用的每种标签的分数。updateTag($user$, $Type$)将返回分数值最高的标签。标签分数的计算公式如下所示：

$$Score(Neighbor.\ Tag[i])=\frac{\sum_{j}^{|Type|}W_{Type[j]}Neighbor_{Type[j]}.\ Tag[i].\ time}{\sum_{j=1}^{|Type|}\sum_{t=1}^{|Tag|}W_{Type[j]}Neighbor_{Type[j]}.\ Tag[t].\ time}$$

式中，$Neighbor.\ Tag[i]$ 表示邻居节点标签集中第 i 种标签，$Neighbor_{Type[j]}.\ Tag[i].\ time$ 表示关系为 $Type[j]$ 的邻居节点使用第 i 种标签的次数，$W_{Type[j]}$ 表示关系 $Type[j]$ 的权值。

标签更新停止后，将得到更新后的 RDF 图。通过如下 SPARQL 查询，可以得到共享同一标签的节点分组：

```
select ? use ? tag from〈G〉where{
  ? use scot:hasTag ? tag
}group by ? tag
```

但是这些节点分组，不一定代表一个社区结构。将上述查询中获取的标签值存入数组 Tag 中，并对 RDF 图作进一步检测，共享同一标签且在同一成分内的节点构成社区结构。假设 Tag 中存在 T 个标签，则通过运行 T 次下述 SPARQL 查询，就能实现社区的发现：

```
select ? x ? y ? tag from〈G〉where{
  ? x scot:hasTag Tag[i];
  ? y scot:hasTag Tag[i];
  ? x Type[rel] ? y
}group by ? x ? y
```

3.2　算法性能分析

设 RDF 图中的用户节点为 n，节点间存在 m 条边，ISLPA 算法中的标签更新的迭代次数为 k，则有 $k<<n<m$。算法对 RDF 图进行语义预处理耗费的时间复杂度为 $O(n)$。每一次迭代更新标签时，需要遍历所有边，耗费的时间复杂度为 $O(m)$。经 k 次迭代完成标签迭代，算法的总时间复杂度为 $O(km)$。迭代结束后，检测共享同一标签的不同成分内的节点构成社区结构，需要消耗的时间复杂度为 $O(n+m)$。算法的总时间复杂度为 $O(2n+(k+2)m)$。由于 $k<<n<m$，因此算法仍接近线性时间复杂度，适用于大规模的在线社会网络的社区发现。

4　实验分析

实验 1　对经典社会网络的分析

为验证算法的有效性，首先使用 ISLPA 算法对经典社会网络图进行分析，并将社区发现结果与传统算法的社区发现结果进行比较。实验选择《悲惨世界》人物关系网络作为分析对象，用复杂网络分析软件 Gephi[9] 对网络基本特征进行分

析，该网络包括有 77 个节点和 254 条边，网络直径为 5，平均最短路径为 2.4，密度为 0.043，聚类系数为 0.287。

使用 ISLPA 算法对网络进行社区划分，算法中用人物名作为节点的初始标签，并将网络中所有边的权值初始为 1。图 5 记录了 3 次运行算法时网络模块度的演化过程。

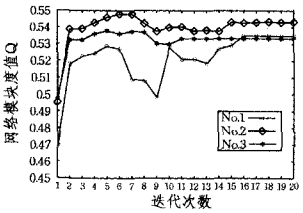

图 5　3 次运行 ISLPA 算法的模块度演化图

3 次运行 ISLPA 算法所得的社区划分结果如表 1 所列。3 次运行算法都将网络分为 5 个社区结构。结合对各个社区中成员的节点度数分析，可以发现各个社区中的核心节点。例如 3 次划分中都存在用“Myriel”标识的社区结构，经过分析可得，Myriel 节点的度数为 10，是该社区中最大的节点度，因此 Myriel 为该社区结构的核心节点。在小说中 Myriel 确实是主要角色之一，这也证明了分析的正确性。

表 1　3 次运行算法所得的社区划分结果

运行次数	划分结果(社区标签及对应的成员数)	模块度值
No. 1	Myriel(10), Blacheville(11), Fantine(32), Chenildieu(6), Mabeuf(18)	0.5348
No. 2	Myriel(8), MmeMagloire(19), Blacheville(10), Javert(22), Enjolras(18)	0.5430
No. 3	Myriel(8), Tholomyes(10), Chenildieu(19), Bossuet(25), Babet(15)	0.5331

此外，本文还使用 CNM 算法[3]、随机游走算法[10]和标签传播算法对《悲惨世界》人物关系网络进行社区划分。各种算法的社区划分结果如表 2 所列。选用 3 次运行 ISLPA 中 Q 值最大的情况与其他算法进行比较。其中，标签传播算法和 ISLPA 算法所得结果的 Q 值略高。表 2 中含有两次运行标签传播算法得到的不同结果，表 2 中用后缀 A 和 B 标注。在不考虑标签的语义关系和网络边的权值差异时，ISLPA 算法和标签传播算法的效率相当。为验证标签语义关系和边的权值对划分结果的影响，在实验 2 中将对真实在线社会网络进行实例分析。

表 2　各种算法的社区划分结果

算法	社区数目	社区成员数	模块度值
CNM 算法	5	{7,13,26,15,6}	0.5005
随机游走算法	8	{16,10,25,8,3,2,7,6}	0.5214
标签传播算法 A	7	{21,10,12,2,6,8,18}	0.5387
标签传播算法 B	6	{10,17,10,11,8,21}	0.5481
ISLPA 算法	5	{8,19,10,22,18}	0.5430

实验 2　在线社会网络的实例分析

为验证算法对真实在线社会网络分析时的有效性和可靠性，设计了一组对比实验，在同一数据集上运行了 3 种社区发现算法，分别是 LPA 算法、SemTagP 算法和 ISLPA 算法。实验数据集取自豆瓣网，通过公开 API 收集用户个人信息、用户关注对象信息以及用户的书籍标签信息。数据集包括了 301 个用户和 2850 个标签，用户间存在 5 种交互关系，这 5 种关系及其对应的权值如表 3 所列。对数据集进行预处理消除噪声信息后，用 Gephi 对网络基本特征进行分析，该网络的直径为 11，平均最短路径为 3.934，密度为 0.035，聚类系数为 0.254。在 2850 个标签中存在 1943 对“skos:narrower”语义关系，不同标签共有 532 种。

表 3　5 种关系对应的权值

关系标识符	关系类型	说明	权值
rel:CloseFriendOf	密友	双向关系	5
rel:WorksWith	同事	双向关系	4
rel:ColleagueOf	同行	双向关系	3
rel:KnowsByReputation	通过知名度认识	单向关系	2
rel:KnowsInPassing	很浅的交情	双向关系	1

实验流程如图 6 所示。LPA 算法是利用网络的结构特征发现社区，所以实验中首先要从原始数据集中提取用户及其关注用户的信息，然后用规范化格式描述这些数据，最后运行算法得到划分结果。而后两种算法在分析中考虑了网络的语义信息，需要对数据集的语义信息进行加工，使用已有本体将网络描述为 RDF 图。

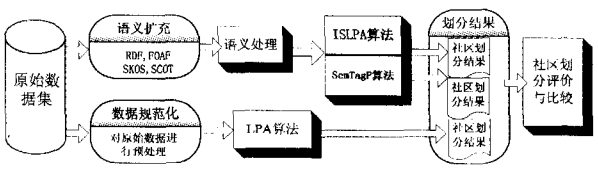

图 6　对比实验流程图

经过实验，得到 3 种算法运行时各个阶段的网络模块度值。模块度的演化过程如图 7 所示。实验中，SemTagP 算法和 ISLPA 算法使用相同的标签语义关系。为验证不同层次的标签语义对社区划分结果的影响，实验中，在两种标签语义关系下使用 SemTagP 和 ISLPA 算法。第一种情况，选择标签层次关系中层次较低的标签作为节点的初始标签时，SemTagP 算法和 ISLPA 算法的运行结果如图 7(a)所示。结果显示，ISLPA 算法获取了较高的模块度，20 次迭代后获取的模块度为 0.683。由于算法 SemTagP 和 ISLPA 都以用户使用的标签作为节点的初始化标签，因此这种算法开始时网络就具有较高的模块度值。SemTagP 算法和 LPA 算法因为均未考虑网络的关系的不同影响力，所以不能获取较高的有向加权网络模块度值。第二种情况，选择标签层次关系中层次较高的标签作为节点的初始标签时，SemTagP 算法和 ISLPA 算法的运行结果如图 7(b)所示。该情况使用的标签产生的社区节点规模较大，与第一种情况相比，SemTagP 算法和 ISLPA 算法划分结果的网络模块度都降低了。由于 SemTagP 划分的情况没有对图 2 的情况进一步处理，导致不合理的大社区结构产生，从而影响了社区划分的效果，结果显示第二种情况下，SemTagP 算法的效果最不理想，模块度仅为 0.352。ISLPA 综合考虑了关系的权值和网络结构特点，且对图 2 的情况做了进一步处理，所以获得了较为理想的划分结果。

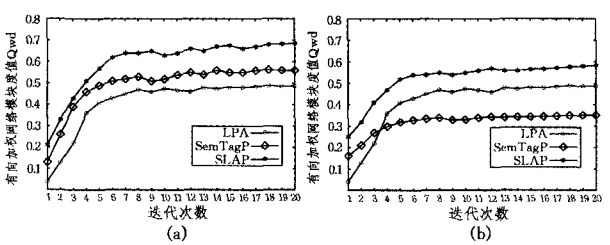

图 7　算法的迭代次数和模块度的演化

第二种情况下，SemTagP 算法和 ISLPA 算法社区划分的部分结果如表 4 所列。SemTagP 算法将社区划分为 13 个，其中部分社区结果内部包括了多个成分。ISLPA 算法在社区划分时也会出现类似情况，但 ISLPA 算法将内部的成分进一步划分，获取了 18 个社区结构。所以 ISLPA 划分结果中含有两个以上标签前缀一样的社区结构，如表 4 中存在两个标签前缀为 programming 的社区结构。

表 4　算法 SemTagP 和 ISLPA 的部分社区划分结果

SemTagP 算法		ISLPA 算法	
标签	用户数目	标签	用户数目
programming	59	novel	44
novel	55	psychology	39
psychology	41	programming_1	33
operatingSystem	33	operatingSystem	32
economic	23	programming_2	25

结束语　针对现有语义标签传播算法 SemTagP 的不足和在线社会网络的特征，本文提出了改进的语义标签传播算法(ISLPA)。ISLPA 算法综合考虑网络中的标签语义关系和用户的各类交互关系，将在线社会网络映射为有向加权网络，并对其进行社区划分。ISLPA 算法接近线性时间复杂度。实验证明，ISLPA 能很好地对真实网络进行社区划分。在下一步工作中，将研究 ISLPA 算法在重叠社区发现上的应用。

参 考 文 献

[1] Newman M E J. Detecting community structure in networks [J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2004, 2(38): 321-330

[2] Kernighan B, Lin S. An efficient Heuristic Procedure for Partitioning Graphs[J]. Bell System Technical Journal, 1970, 2(49): 291-307

[3] Clauset A M, Newman M E J, Moore C. Finding community structure in very large networks[J]. Physical Review E, 2004, 6(70): 1-6

[4] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E, 2003, 6(69): 5-11

[5] Palla G, Derényi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society [J]. Nature, 2005, 7043(435): 814-818

[6] Raghavan U N, Kumara S. Near linear time algorithm to detect community structures in large scale networks[J]. Physical Review E, 2007, 3(76): 106- 115

[7] Ereteo G, Gandon F, Buffa M. SemTagP: Semantic community detection in folksonomies [C] // IEEEWICACM International Conferences on Web Intelligence and Intelligent Agent Technology. 2011: 324-331

[8] 王延鹏. 复杂网络重叠社区发现算法研究[D]. 太原：太原理工大学，2011：24-26

[9] Bastian M, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks[J]. American Journal of Sociology, 2009, 2: 361-362

[10] Pons P, Latapy M. Computing communities in large networks using random walks[J]. Journal of Graph Algorithms and Applications, 2005, 2(10): 191-218

[11] Mika P. Social Network and the Sematic Web [M]. Semantic Web and Beyond, 2007

[12] Passant A, Laublet P. Meaning of A Tag: A Collaborative Approach to Bridge the Gap Between Tagging andLinked Data[J]. Evolution, 2008, 5(41): 1-5

(上接第 19 页)

行开销较小。从图 4 还可以看出，并行加速比有较为稳定的趋势，而优化加速比抖动较大。这是因为算法的优化过程改变了算法的基本数据结构和执行流程，而并行化的过程不改变算法自身结构。

结束语　本文基于 $Z_{3\text{-}2}$ 算法提出了改进的并行凸壳算法。首先，基于颜氏距离来分类点集和找下一个凸壳点，效率优于使用欧氏距离和正负划分；其次，以较小的并行开销实现多核架构下的并行计算；再次，利用阈值判断，避免了深层递归和小粒度任务的分解；最后，从点集的内部开始，逐步迅速地删除大量的非凸壳点，避免了大量的无效重复计算。理论和实验结果表明，改进算法利用颜氏距离减少了计算量，在多核架构下充分利用并行计算资源对海量数据集进行求解，并行开销小，获得了比原 $Z_{3\text{-}2}$ 算法更高的效率。因此，算法不仅能求解一般性凸壳问题，而且能更快速地求得海量(10^6 个以上)的点集和含有大量非凸壳点的点集的凸壳。本文提出的理论和思想可以应用到三维和更多维的情况中，这也是下一步的研究目标。

参 考 文 献

[1] 周培德. 计算几何-算法分析与设计(3 版)[M]. 北京：清华大学出版社，2008

[2] Graham R L. An efficient algorithm for determining the convex hull of a finite planar set[J]. Information Processing Letters, 1972, 1(1): 132-133

[3] 金文华，何涛，刘晓平，等. 基于有序简单多边形的平面点集凸包快速求取算法[J]. 计算机学报，1998，21(6)：533-539

[4] 赵军，曲仕茹. 平面点集凸壳的快速算法[J]. 计算机工程与应用，2009，45(1)：56-58

[5] Reif J H, Sen S. Optimal parallel randomized algorithms for tree-dimensional convex hulls and related problems [J]. SIAM J. Comput., 1992, 3(21): 466-485

[6] 张三元，马利庄. 平面散乱点集凸包并行算法[J]. 浙江大学学报：工学版，1999，33(4)：432-440

[7] 郝小柱，胡祥云，戴光明，等. 平面点集凸包的并行算法研究[J]. 计算机应用，2005，25(10)：2462-2464

[8] 周启海，黄涛，林珣，等. 基于动态基线倾角与基线距离最大化的凸壳并行新算法[J]. 计算机科学，2008，35(4)：244-247

[9] 陈伟，杜凌霞，陈红. 多核架构下的数据处理算法优化策略综述[J]. 计算机科学与探索，2011，5(12)：1057-1075

[10] 邓亚丹，景宁，熊伟. 基于共享 Cache 多核处理器的 Hash 连接优化[J]. 软件学报，2010，21(6)：1220-1232

[11] 张忠武，吴信才. 海量数据凸壳快速优化算法研究[J]. 微计算机信息，2011，27(8)：194-196

[12] 张思乾，程果，陈荦，等. 多核环境下边缘提取并行算法研究[J]. 计算机科学，2012，39(1)：295-298