HipHop算法
利用微博互动关系挖掘社交圈

文 / 张俊林

HipHop算法成功弥补了原有社交圈挖掘算法的不足，能更精准地挖掘微博互动关系中有价值的信息。本文讲述了HipHop算法出现的缘由及价值，同时结合示例给出了算法的实现步骤。

在微博环境下，如何自动挖掘某个用户的社交圈或兴趣圈是个很基础但很重要的问题。准确挖掘某个用户在微博上体现的社交关系，对很多具体应用来说都很有价值。例如，可以更好地对用户的兴趣进行挖掘或者能够推荐用户还未关注的社交圈子成员等，或者根据其社交圈子更准确地对用户进行个性化建模，为其他基于用户个性化模型的推荐或者广告推送等提供基础服务。

我们在微博相关研发任务中提出了HipHop算法，旨在通过利用微博用户的互动行为，来自动挖掘出用户的不同社交圈子。在设计算法之初，我们希望圈子挖掘算法能同时满足以下几个条件。

- 对于某个微博用户A来说，可以挖掘出其所属的多种社交圈子，例如用户既有的同事关系圈、所属的专业兴趣圈等。
- 同时对于用户B来说，他可能同时属于用户A的不同社交圈。比如B既是A的大学同学，也是A的某公司同事，那么B应该同时出现在用户A的两个不同兴趣圈里。
- 不使用用户隐私数据。出于保护用户隐私的目的，我们希望算法只使用用户的公开行为和信息，因此HipHop算法只使用了互动关系这种公众完全可见的公开信息。
- 社交圈可解释，即可以通过简洁的方式描述社交圈的性质或者特点。目前是通过给每个圈子打上不同的标签来进行区分。

HipHop社交圈挖掘算法就是根据这几个指导原则设计开发出来的，它能够同时满足以上几个约束条件。目前公开的参考文献中很少见到能够同时满足这些条件的相关社交圈挖掘算法。

常见的社交圈挖掘算法

社交圈挖掘是目前社交网络研究中非常典型和热门的研究任务，通常被称为“社群发现”。学术界也陆续提出了很多算法来解决这个问题。大体而言，可以将其分为两大类：“单社群”方法和“多社群”方法。所谓“单社群”方法，就是说网络结构中的某个节点只能隶属于某个社群，不允许出现隶属多个社群的现象。而“多社群”方法则允许用户同时隶属于多个社群。下面分别以GN算法和“最大团结构”作为这两类算法的代表对其思路进行简要介绍。

GN算法

GN算法是一种非常常用的图结构中社群自动发现

算法，最初由Girvan和Newman在2002年提出，因其有效性得到了广泛的使用。

GN算法的基本思想是：在图结构中，首先计算每条边的“介数”，然后从图中删除“介数”最大的边，如此不断循环，一直迭代删除当前“介数”最大的边，最终就形成了发现的社群。所谓边的“介数”，指的是图中任意两个节点的最短路径中经过这条边的次数。边的“介数”越大，则这条边是连接了两个或者多个社群或者圈子的多余边的概率越大，因此通过不断删除高“介数”边可以达到分离社群的目的。

GN算法比较有效，但它是一种“单社群”发现方法。就是说，图中某个节点只能属于固定的一个社群，不可能同时属于多个社群。这与实际应用场景需求有较大差异，因此成了该算法的局限。

“最大团结构”算法

“最大团结构”（max clique）是一种比较流行的能够进行“多社群”发现的算法，即图中的节点可以同时隶属于多个社群。

“最大团结构”能对图的拓扑结构进行分析，找到满足“最大团”性质的子图结构，即最大的全联通子图，每个“最大团”就是一个发现的社群。

尽管“最大团结构”算法可以发现某个节点属于多个社群，比“单社群”发现方法有更多的实用性和应用场景，但这个算法也有其局限：因为“最大团结构”要求是全联通子图，即子图中任意两个节点都有边连接，这是一种非常强的约束。真实应用中能满足如此强约束的图结构很少，这导致这个算法很多图中的节点无法归入某个社群。

HipHop算法的某些步骤中也采取了“最大团结构”的思想，但通过技术手段放松了这种约束，有效地改进了其效果。

利用HipHop算法发现社交圈

Hiphop算法利用微博用户的互动关系来自动挖掘某个用户的不同社交圈。这里的“互动”是一种总称，包括转发微博、评论微博和@其他用户等行为。如果用户A和用户B有任意上述提到的行为，则可以认为两者间有互动关系存在，且根据其频率可以赋予边不同的强度，代表了两个用户的社交亲密程度。

我们之所以使用社交关系来挖掘社交圈，是基于以下的一个基本假设：与某个微博用户进行过交互行为的人群存在于不同的小团体中，而小团体成员之内有较为密切的互动行为，不同小团体之间、成员之间交互行为较少。比如你的大学同学之间在微博上有较多互动行为，但他们和你的同事之间就很少有交互行为（如图1所示）。尽管这只是一种假设，但实际挖掘效果表明大多数情况下这种假设是成立的。

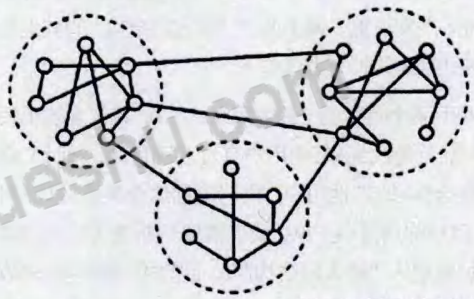

图1 不同团体的社交行为

HipHop算法的技术流程可以划分为顺序进行的三个步骤。

步骤一：从与用户有直接互动的其他用户中寻找“最大团结构”。

首先，对于微博用户A，所有与用户A在微博上有过直接互动行为的用户形成直接互动集合S。本步骤试图在集合S中找到多个“最大团结构”，即挖掘多个小团体的核心成员。

对于集合S中的节点来说，可以根据它们相互之间的互动关系构造一个图G，并在此基础上去挖掘图G中的“最大团结构”。所谓“团结构”，就是图G中包含的任意全连通子图。例如，图G中的三个节点{a,b,c}，如果它们之间任意两人都有互动关系存在，则形成了一个三节点的“团结构”。

如果某个团结构能够纳入新的节点形成新的团结构，那么它就不是最大的，最大团结构是不能容纳其他更多节点的团结构。比如上述的三节点团结构，如果存在节点d，这个节点和a、b以及c都有互动关系，那么{a,b,c,d}就形成了一个四节点的“团结构”，而如果找不到节点能够和{a,b,c}都有

互动关系，那么{a,b,c}就是一个三节点的“最大团结构”。

图的“团结构”是一个非常强的约束，因为它要求图中任意两个节点都存在互动关系。步骤一找出的某个用户A的“最大团结构”的物理含义是：与用户A有密切关系的那些用户中，有哪些是有密切联系的小团体。

步骤二：“最大团结构”在直接互动用户集合中的扩充。

步骤一找出了与用户A有过直接互动行为的集合S中形成的“最大团结构”。步骤二在此基础上，在集合S范围内对每个发现的“最大团结构”进行扩充，来发现更多属于某个“最大团结构”的其他用户。具体的扩充方式如下。

对于某个具体的“最大团结构”T，其包含若干用户，首先找到和T中用户有过互动行为，同时又在集合S中的其他用户，我们简称这个集合为U。对于U中的某个用户w，我们需要判断是否应该将其扩充进入“最大团结构”T。目前的判断标准采取如下公式：

$$Utility(G)=\frac{\sum_{a\exists E_{in}} Weight(E_a)}{\sum_{b\exists E_{out}} Weight(E_b)}$$

假设G是最大团T将用户w融合后形成的新图，公式的分子部分代表新图G中所有节点内部边的权重之和，而分母部分代表图G中所有节点和图G之外的任意节点形成的所有边权重之和。如果Utility(G)函数比未扩充节点w的原图结构T的效用函数Utility(T)值大，那么认为将节点w扩充进入T是合理的，否则不能将节点w扩充进入图T中。有这个函数作为标准，我们就知道集合U中的用户哪些该扩充进入团结构T中，哪些该被舍弃。

之所以采取上述公式作为判断标准，是基于之前提到的如下假设：一个社交圈子成员之间互动关系密切，而圈子成员与圈子外成员之间的互动关系不是很密切。上述公式就是这个基本假设的具体体现，分子部分是衡量圈子成员内部的关系紧密程度，而分母衡量的是圈子成员和圈子外成员的关系紧密程度。

从公式可以看出，如果圈子成员之间互动越多，而与圈子外成员互动越少，则效用函数越大，也就是说这个圈子越紧密。

如果对于集合U中所有后续扩充用户都采用上述公式进行判断取舍，来做出是否将这个用户扩充进入“最大团结构”T的决策，那么就完成了T的一轮扩充，形成了扩充后的新集合T'。对于T'来说，仍然可以采取上述扩充方法不断外扩。“最大团结构”T外扩的终止条件是：如果对于集合U中所有用户，做出的决策都是不进行扩充的，那么此时已经达到了扩充的边界，可以停止外扩，形成最终扩充结果。

如果对步骤一中发现的所有“最大团结构”都采取上述方式外扩，就完成了步骤二的任务。可从上述过程看出，步骤二是对步骤一的扩充阶段。

步骤三：与用户有“二级互动”关系的其他用户集合中的扩充。

所谓用户A的“二级互动”用户集合，是指与用户A有直接互动的用户形成集合S，而与集合S中任意一个用户有互动行为的所有其他用户形成了二级互动集合。

从步骤二的结果来看，步骤二完成了对“最大团结构”的扩充，在直接互动用户集合中找到了不同的社交圈子。步骤三首先将直接互动用户集合S扩充为二级互动用户集合，然后采取与步骤二类似的方法继续向外扩充，这样就形成了HipHop算法的最终结果，形成了用户A的多个不同社交圈，而任意一个其他用户B可能同时属于用户A的多个社交圈。

通过上述三个步骤，就可以利用微博互动关系自动挖掘出某个用户的社交关系圈。对于微博的海量用户而言，只要对每个用户都依次采取上述步骤，即可获得最终结果。这可以采取大规模并行计算来快速实现。

下面结合一个例子讲述HipHop算法。以“李开复”作为示例，说明上述步骤及其中间输出结果。

对于步骤一，首先找到与“李开复”有过互动的微博成员形成集合S，之后在集合S里采用发现“最大团结构”的方法，可以得到最初的5个“最大团结构”：

- 最大团1（创新工场有关）：王肇辉/蔡学镛/周源/张亮/徐磊Ryan
- 最大团2（互联网媒体相关）：keso已被XX/牛立雄/金磊
- 最大团3（财经投资相关）：徐小平/爱国者冯军/潘石屹/杨澜
- 最大团4（创新工场有关）：郎春辉/罗川/袁聪iw/应用汇
- 最大团5（企业家相关）：曹国伟/江南春/吴征bruno/蒋锡培

经过步骤二，对原始的5个最大团在集合S中进行扩充，每个原始的最大团都有不同程度地扩大，其新扩充进的成员范围在3~10个。

步骤三，首先将直接互动成员集合S扩充为二级互动成员集合，即将与集合S中成员有过互动行为的微博用户形成新的更大范围的集合。根据前面讲述的扩充方式，5个最初的“最大团结构”获得进一步扩充，最后形成了多个不同的社交圈。

经过人工评估，HipHop算法挖掘出的社交圈有较强的社交内聚度，同时也满足算法设计之初设定的几个约束条件，因此具有很强的实用性。同时，经过大量实例分析，我们发现在微博中形成的社交关系和IM形成的社交关系有较大的差异，大部分用户的微博中的社交关系以同事关系和兴趣关系为主，而IM中形成的社交关系则以亲友、同事、同学等线下关系为主，这可能反映了社会化媒体和传统社交网络的区别所在。

张俊林

《这就是搜索引擎：核心技术详解》作者，新浪微博研发人员，中科院软件所博士，主要研究方向为自然语言处理、搜索技术、推荐系统及机器学习。

责任编辑：杨爽（yangshuang@csdn.net）