## **1.** **Personalized Influential Topic Search via Social Network Summarization（7.14李昊展）**

通过社交网络摘要进行个性化的有影响力的主题搜索

### 1.1 基本信息

期刊：《IEEE Transactions on Knowledge and Data Engineering》

发表时间：2016

作者：Jianxin Li、Chengfei Liu、Jeffrey Xu Yu、Yi Chen、Timos Sellis、J. Shane Culpepper

作者信息：

School of Computer Science and Information Technology, RMIT, Australia.

Faculty of Science, Engineering and Technology, Swinburne University of Technology, Australia

Department of Systems Engineering & Engineering Management, The Chinese University of Hong Kong

College of Computing Sciences, New Jersey Institute of Technology, USA

Faculty of Science, Engineering and Technology, Swinburne University of Technology, Australia

School of Computer Science and Information Technology, RMIT, Australia.



### 1.2 文章主要研究内容

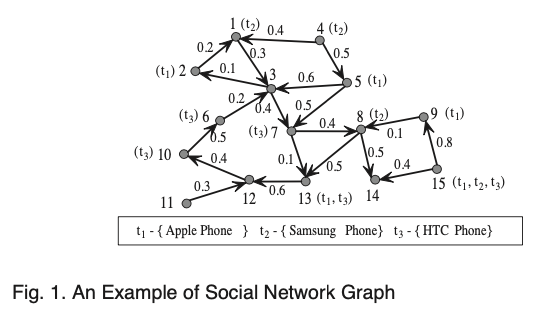
社交网络是向朋友和同事传播信息的重要机制。在这项工作中，我们调查了一个重要问题——社交网络中的个性化有影响力的主题（PIT）搜索：给定用户u在社交网络中发布的关键字查询q，PIT搜索是找到对查询用户u最具影响力的顶级q相关主题。主题对查询用户的影响取决于**查询用户与社交网络中包含该主题的社交用户之间的社交联系。**为了在类似的粒度尺度上衡量主题的影响，我们需要提取社交网络关于主题的社会总结。为了进行有效的主题感知社会总结，我们提出了两种基于随机行走的方法：随机聚类和L长随机行走。根据所提出的方法，我们可以找到一小部分具有影响力分数的代表用户，以模拟社交网络中大量主题用户对该主题的影响。选定的代表用户表示为在社交网络上传播的主题感知影响的社会总结。然后，我们通过将其应用于个性化有影响力的主题搜索问题来验证社会总结的有用性。最后，我们使用现实世界的数据集评估了算法的性能，并表明该方法在实践中是高效和有效的。

### 1.3 文章内容

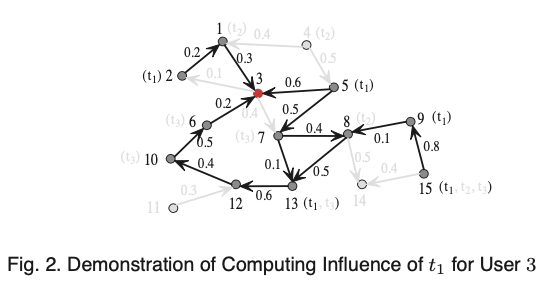
**1.问题提出**

Twitter、Facebook和微信等社交网络为用户分享信息提供了便利的平台，其重要性不断增强。近年来，这些网络中的信息和用户连接的动态性质提出了许多有趣的开放性研究问题，如影响力最大化问题，以及话题检测问题。然而，一个发布关键词查询的用户很容易被查询相关的话题数量所淹没。这是因为一个大的社会网络可能包含数百万的社会用户分享对各种事件或话题的评论，而这些评论又会导致新的话题。在这种情况下，识别一小部分与查询相关的话题是一个具有挑战性的问题。最广为接受的方法是**根据话题和查询之间的术语相关性来选择相关的话题**，其方式类似于典型的关键词搜索。另一种方法是**为查询推荐最新的查询相关话题**，但这在实际应用中可能并不适合，因为它没有考虑较早的话题或讨论。为了解决这个问题，我们希望**选择一组能够充分总结给定话题的有代表性的用户，其中代表性的用户是根据他们与话题用户的接近程度来评估的**，这些用户发布的信息包含了话题术语。在这种情况下，如果查询用户有可能被选定的代表用户所接触或影响，那么查询用户就有很大可能关注或信任选定的代表用户所推荐的话题。在本文中，我们把这个问题称为个性化影响性话题搜索，或者更简洁地称为PIT-搜索。我们的目标不是要找到一个话题如何影响一群用户，而是要**找到如何更好地利用重要的话题和有影响力的用户来满足特定用户的信息需求**。换句话说，如果用户在社交网络中处于类似的社会环境中，当他们发出相同的关键词查询时，他们会看到相同的结果。PIT-搜索可以直接应用于社交网络中的许多个性化服务，如个性化推荐和搜索、目标广告或个人产品推广。

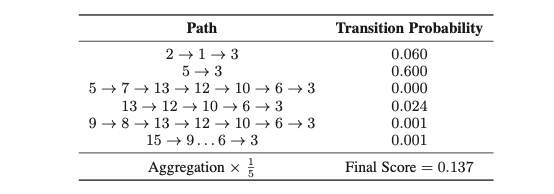
**例子：**考虑图1中描述的小型社交网络，其中一个节点代表一个用户，边的权重代表用户对邻居的影响。边缘的方向显示了两个节点之间的影响关系【有向加权网络】。这里，假设用户对正在讨论的与手机有关的话题表达了积极的意见，如苹果手机为t1，三星手机为t2，HTC手机为t3。一个用户可能会像用户13那样提到几个不同的手机。



现在假设用户3想通过查询q = {Phone}来查看她的社交网络，从而知道哪款手机是她的最佳选择。在这种情况下，三款手机t1、t2和t3都是与q相关的话题，而且有三个邻居用户（用户1-t2，5-t1，6-t3）可以影响用户3，使选择变得困难。



我们对这一挑战的解决方法如下。给出图2中的话题t1，影响过渡路径和相应的过渡概率如下所示。由于包含t1的用户有五个，所以假设每个节点都有相等的本地权重1/5。一个话题对特定节点的最终影响力得分可以通过本地权重和过渡概率相乘来计算。



如示例1所示，PIT-Search可以为不同的用户返回不同的主题结果，即使他们发出相同的查询。在进行高效和有效的PIT搜索时，有三个问题需要解决。（1）社交网络可能有大量与q相关的主题。（2）对于每个q相关主题，PIT搜索必须计算更多主题相关用户对给定用户和查询的影响。（3）任何给定的搜索都可能被几个有影响力的邻居错误地主导。为了减少这种偏见，我们需要通过选择具有代表性的用户来为社交网络推导出主题驱动的社会总结。

为了解决上述具有挑战性的问题，我们通过开发以下三种技术来研究PIT-搜索的问题：**话题敏感的代表节点选择、话题到代表的用户指数和个性化传播指数。**为了进行话题敏感的代表节点选择，并建立话题到代表用户的指数，我们开发了两种新的基于随机漫步的方法，可以根据社交网络中话题的影响传播情况，**只使用话题敏感的代表节点子集来代表所有的话题节点。**一旦我们确定了话题节点和代表节点之间的关系，相应的话题到代表用户的指数就可以随之建立，以捕捉社交网络中小社区的话题敏感的局部影响力。然后，我们构建了一个个性化的传播指数，它构建了附近节点的子集和图中每个节点的相关传播概率。个性化的传播指数与主题和查询都无关。在这两个索引的支持下，我们可以有效地评估任何与q有关的话题的影响力。为了有效地找到对用户最有影响力的前k个q主题，我们还提出了一个动态的top-k PIT-Search算法，该算法基于中间结果来识别top-k结果集中的高质量q相关主题。通过探测尽可能少的节点，将低质量的话题从结果集中修剪出来。利用这三种技术，我们可以大大加快PIT-搜索在大型社交网络中的性能。

**主要贡献：**

- 开发了两种新的基于随机漫步的方法来选择社会网络中的话题敏感的代表节点。这可以捕捉到一个话题在社会网络的不同部分的局部影响。

- 设计了一个top-k PIT-Search算法，根据提出的预先计算的指数，有效地计算出一个用户的k个最有影响力的q相关话题。

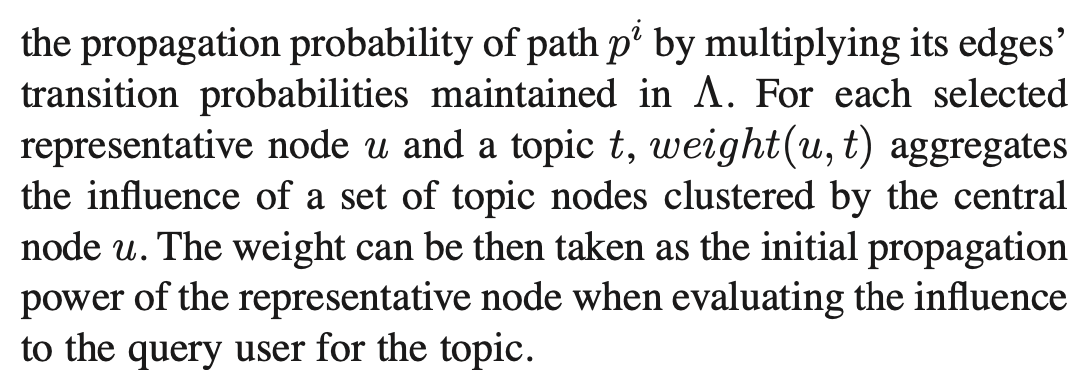
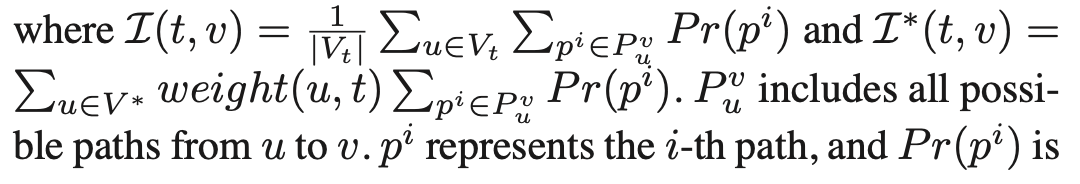
- 使用一个大型的Twitter数据集来评估所提出的技术的效率和效果，并通过与以前的社交网络工作中的三个基线进行比较来显示我们方法的优势。

**2.问题描述**

考虑一个社交网络图G=（V，E，T，Λ），其中V是代表社交网络中的社交用户的节点集，E是描述用户之间联系的边集。T是一个话题空间，每个社交用户可能有一组从用户发布的信息中提取的话题，例如，T（v）={t1，...，}为一个节点v。Λ是E中边的过渡概率。

**（1）社交总结**

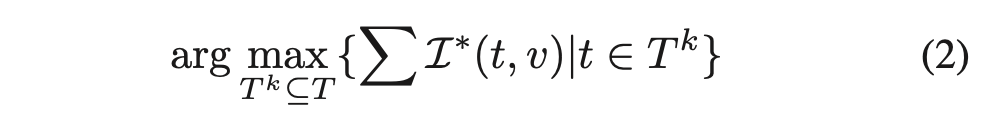
考虑一个社会网络G=（V,E,T,Λ）。给定一个话题t∈T，G的面向t的社会总结是从V中选择指定数量的节点V∗，这些节点代表满足以下条件的话题节点Vt的影响力。



**目的：**使得挑选出的代表节点真的具有话题代表性

**（2）PIT-搜索**

给定一个由用户v发出的关键词查询q，一个社交网络G，以及要返回的项目数量k，一个PIT搜索通过满足的方式返回对v来说最有影响力的q相关的k个话题。



其中，L∗是Tk中的话题对v的总影响力，Tk是一个k大小的话题子集，从T中选择q相关的话题。

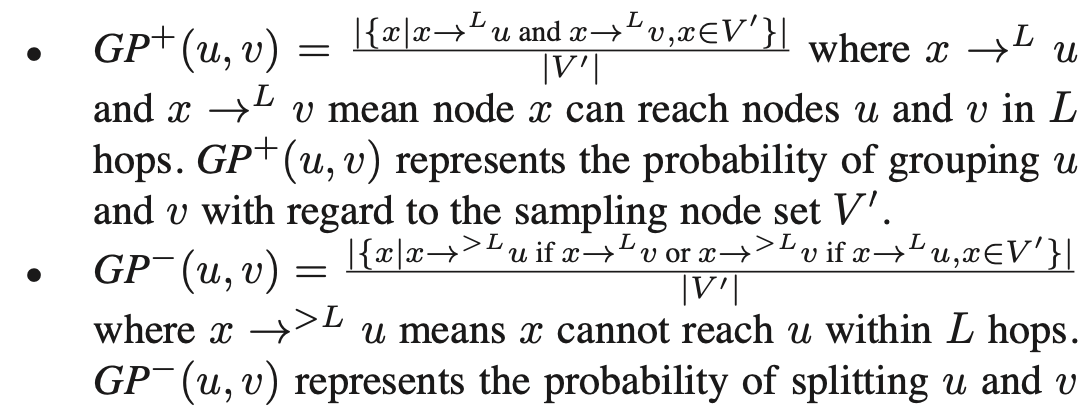
**3.随机聚类（RCL-A）**

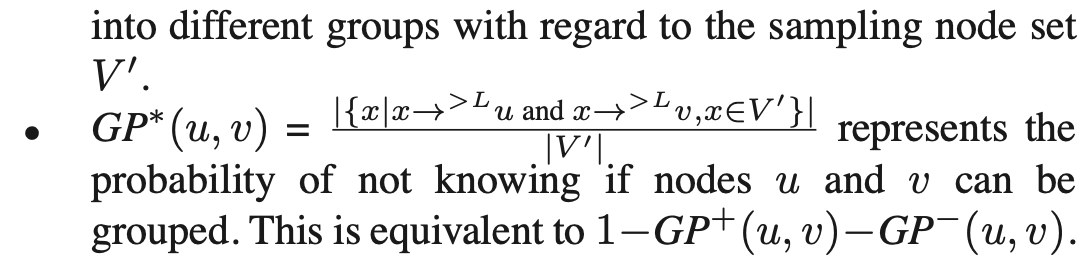
为了选择一组有代表性的话题节点，一个近似的解决方案是对话题节点进行分组，然后为每组确定一个中心节点。最后，所有未被选中的主题节点的本地影响被迁移到中心节点上。然后，中心节点被作为代表节点集。每个代表节点的汇总的本地影响力得分可以用来评估话题对每个社会用户的影响力。

**（1）聚类话题节点**

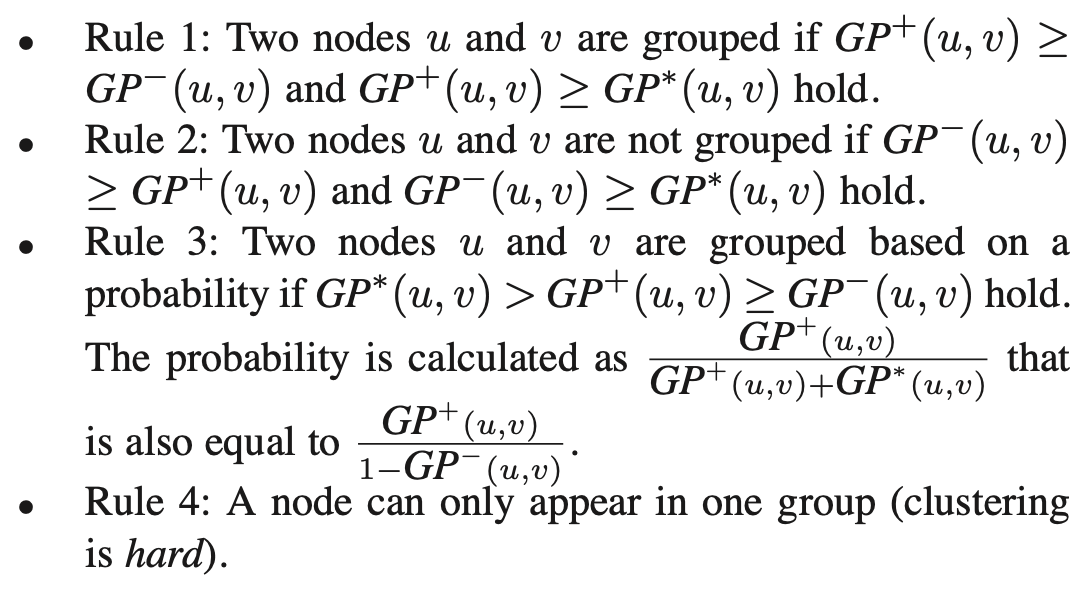
对主题节点进行聚类的主要思路是，首先测量给定样本节点集的主题节点的**共同可达性**，然后根据共同可达性将主题节点归入不同的聚类。这里，一个**节点的可达性是指该节点可以到达的样本节点的数量**。给定两个节点，共同可达性被评估为两个节点共同到达的样本节点的数量。基本假设是，如果两个节点可以被归入一个群组，那么这两个节点很有可能可以到达一定数量的共同邻居。换句话说，**节点能达到的共同邻居越多，将这两个节点分组的概率就越高。**

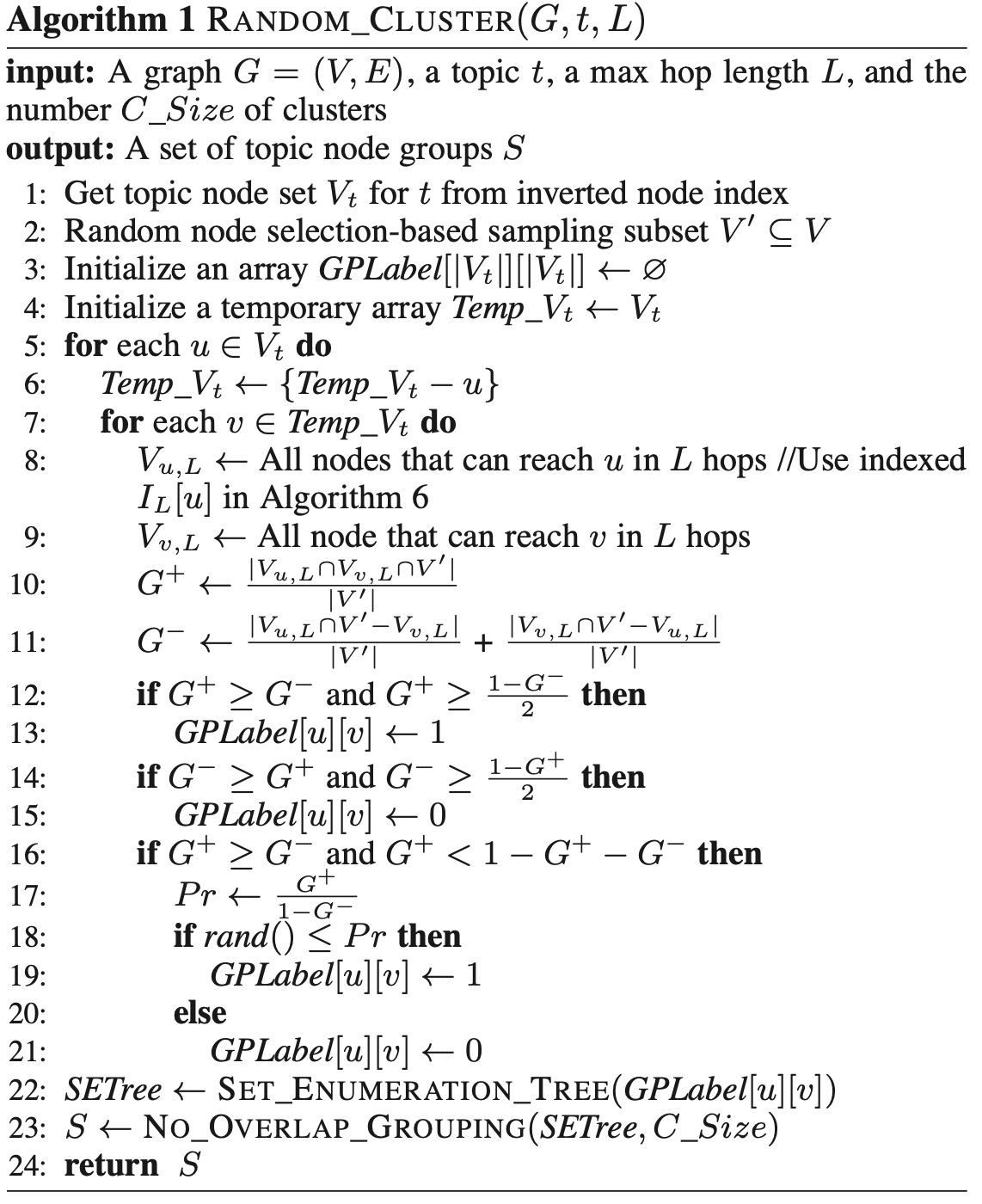
为了决定两个节点是否可以被分组，RCL-A的解决方案是基于以下概念。首先，从V中选择一个采样节点集V′**（基于节点度值为概率进行采样）**。用三种变量来衡量两个节点u和v的分组关系。





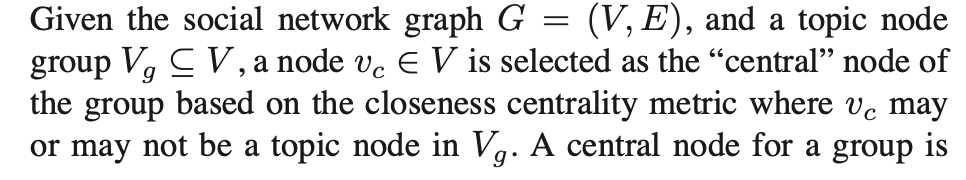
**（2）聚类规则**

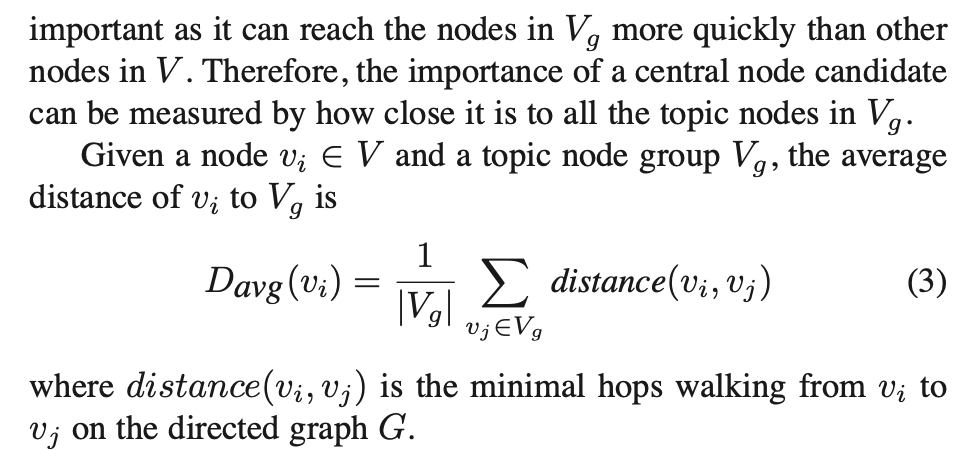
****

****

**（3）挑选中心节点**

得到不同的组别后，再基于接近中心性，挑选出网络的各组的中心节点，使得中心节点可以更快到达Vg中的节点

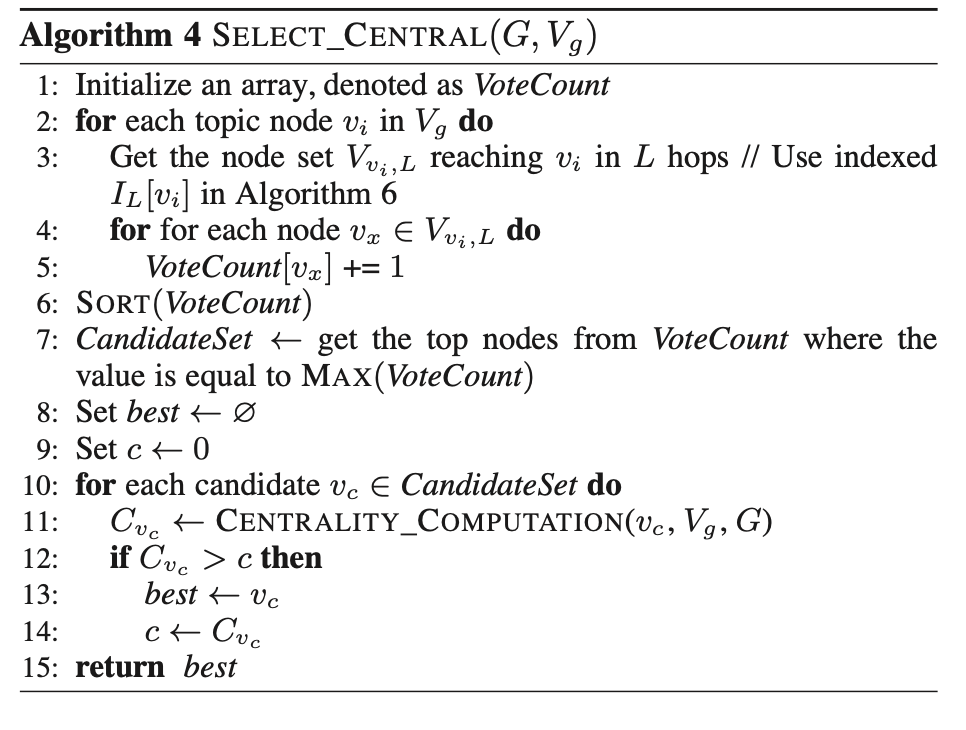
****

****

尽管节点的亲密度中心性等同于所有对最短路径问题，但由于该问题的精确解决需要花费Θ(|V |3)，所以处理大型图并不实际。本文目标是选择一个靠近给定主题节点组的节点，并且该组中任何两个节点的最大距离被限制为2L。因此，V中的许多节点可以从计算中修剪掉。为了有效地找到一组主题节点的中心节点，本文提出了一种启发式的方法，首先不使用CC来识别几个中心候选人，然后通过计算候选人子集的接近中心度来找到最佳的中心节点。

**思想：基于Vote方法缩小候选集，再基于接近中心性挑选中心节点**

首先计算Vg中的主题节点在L跳内到达的共同节点，并存储能够到达共同节点的主题的投票数（使用VoteCount）。换句话说，VoteCount代表共同节点的命中频率。在VoteCount中得票最多的节点成为候选集。对于每个候选者，我们用CC计算出接近中心度。



**局限性：**

（1）限制一个群体产生的中心节点可能会增加大群体中的中心节点和小群体中的中心节点之间的影响力偏差。这可能会导致中心节点高估或低估社会网络中的本地影响力。

（2）在将影响力分配给代表节点时，将每个小组限制在一个话题节点可能并不精确。这是因为一个话题可能有不同的代表，有不同的概率。

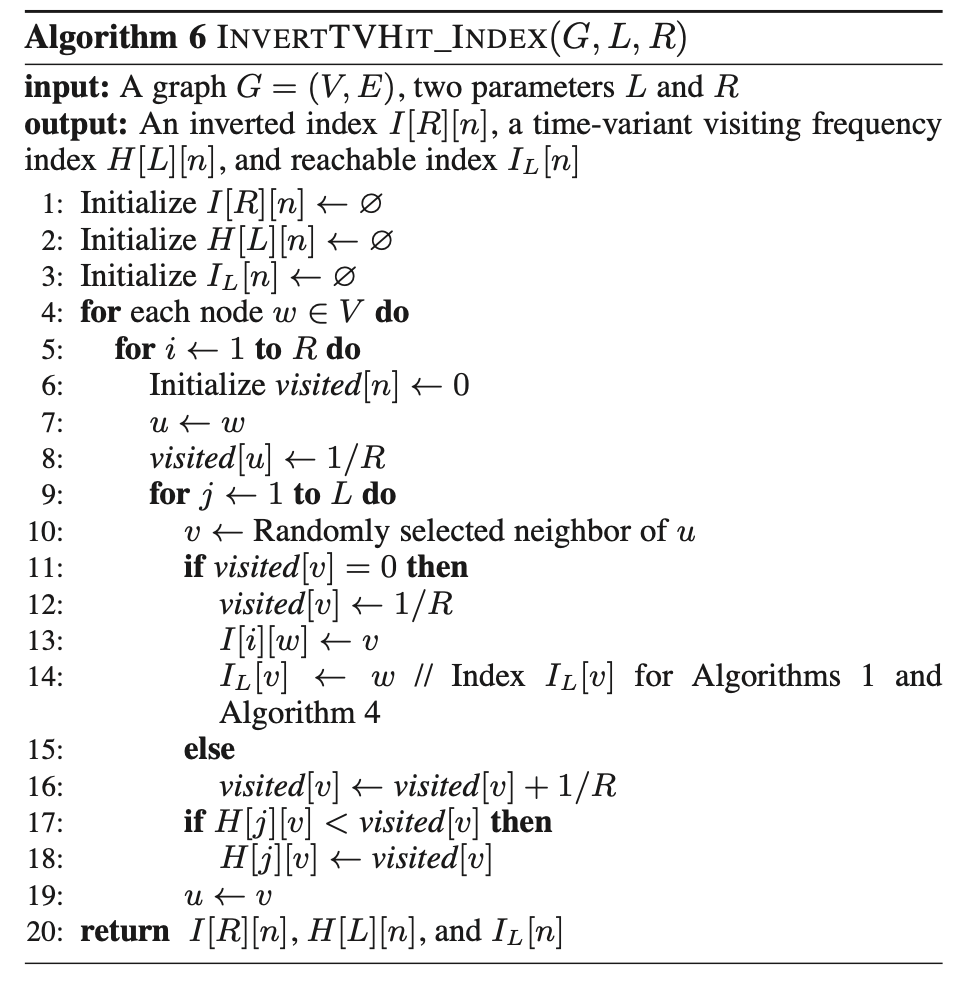
（3）RCL-A是基于一个时间的聚类方法。在确定给定主题节点的中心点时，只考虑 "重要 "路径可能是不够的。这是因为一些经常被访问的路径在评估节点的中心性时将发挥重要作用。

（4） 用RCL-A计算中心节点是一个昂贵的计算，而且生成的组的数量可能非常大。

**4.L步长随机游走（LRW-A）**

为了解决RCL-A的局限性，提出了一种新的方法，即基于L长度的随机漫步技术，为主题选择代表性节点。首先，我们为每个节点取R个L长度的随机行走的样本。然后，预先计算出的数值可以用来降低为社会网络中的任何话题生成代表节点集的成本。最后，话题节点的局部影响可以使用吸收性随机漫步迁移到代表节点上。

**（1）反向索引【挑选R个L长度都随机行走样本】**

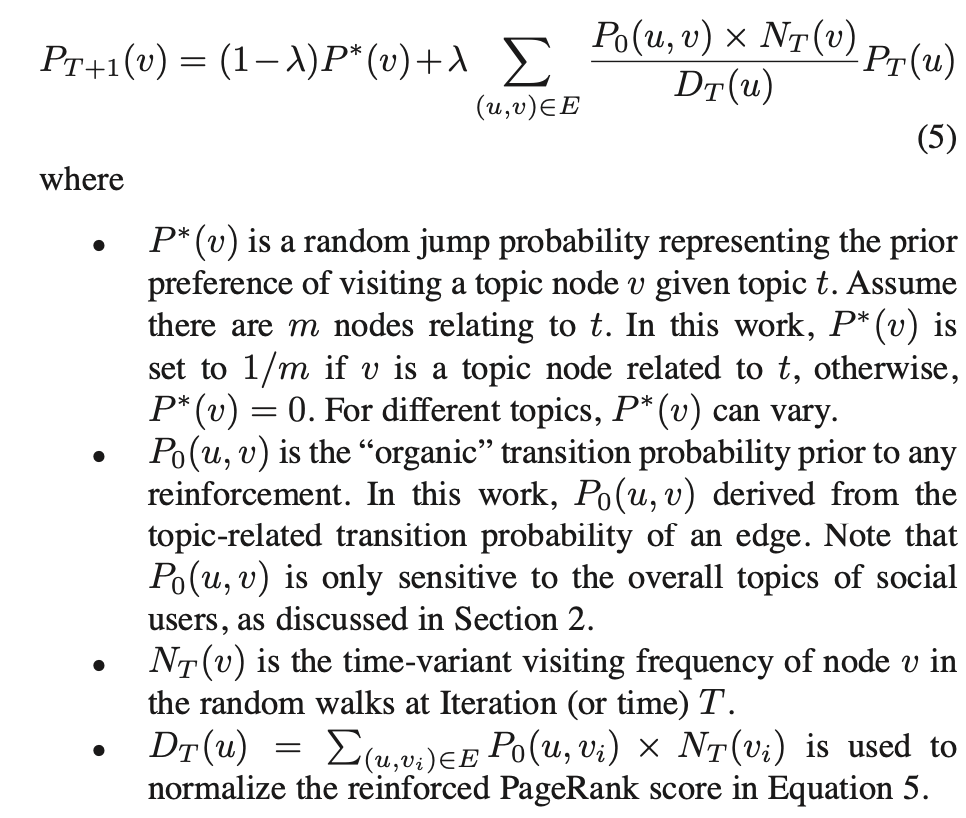


为了索引随机行走样本，构建了R个反向的列表，每个列表包含n个子列表。对于每个节点u，一个子列表包含了从节点u出发的L长度的随机游走。此外，一个时间变化的访问频率指数H[L][n]被构建，以衡量一个节点在L次迭代中被访问的频率。

算法6描述了在G上执行基于样本的随机行走时如何构建索引。R反向列表，用I[R][n]表示，其中I[i][w]索引从节点w开始的第i个L长度的随机行走。这里H[L][n]被用来保持每个节点在[1，L]时间段的时间变化访问频率，即Iteration-1到Iteration-L。首先，该算法初始化数组I[R][n]和H[L][n]。然后，对于V中的每个节点w，算法从节点w开始执行R个L长度的随机行走。对于每个L长度的随机行走，路径的选择是通过随机选择节点u的一个邻居（在第一次迭代中u=w），然后迭代地用选择的邻居v替换节点u。

**（2）有效的代表节点筛选【挑选代表节点】**

给定G中与主题t相关的节点集Vt，本小节的任务是选择一组节点，成为节点集V的代表节点。所选择的节点集应该是一个接近最优的节点集，并尽可能接近Vt。为此，本文开发了一个类似于PageRank的新型个性化模型，根据中心性和多样性对节点进行排名。



P∗(v)是一个随机跳转概率，表示访问给定主题t的主题节点v的优先级。假设有m个节点与t有关。在这项工作中，如果v是与t相关的主题节点，则P∗(v)设置为1/m，否则，P ∗(v) = 0。对于不同的主题，P ∗(v)可以有所不同。

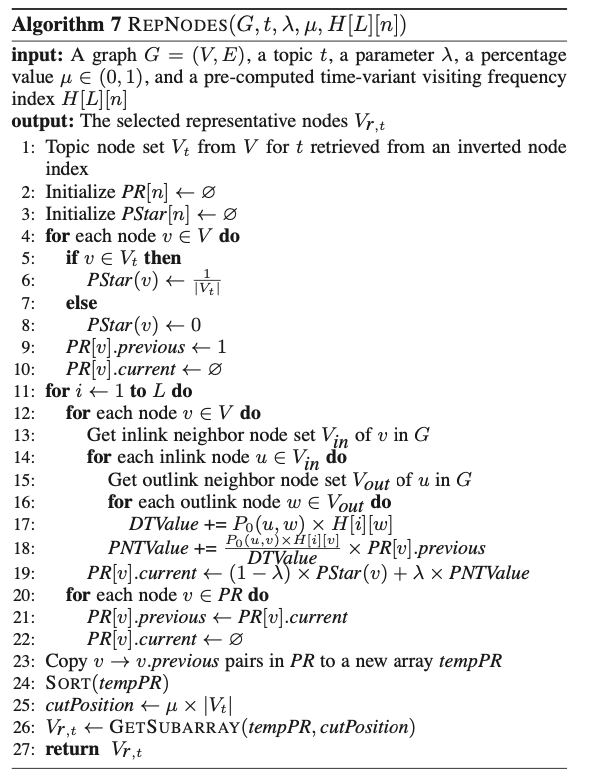
P0（u，v）是初始过渡概率。P0(u,v)来自节点边与主题相关的过渡概率。

NT（v）是迭代（或时间）T随机游走中节点v的时变访问频率。

用于归一化。

**主要思想**

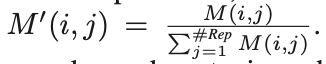
根据**时间变化的点击次数和与主题节点Vt的接近程度**对节点进行排名。由于一个节点的影响力受距离的限制，即L-hops，PageRank式的算法只需要使用L-iterations（T=L）来运行。通过这样做，每个节点的权重都是基于L长度半径内的节点。



**（3）主题节点的局部影响力迁移【找最近的代表节点，将影响力转移】**

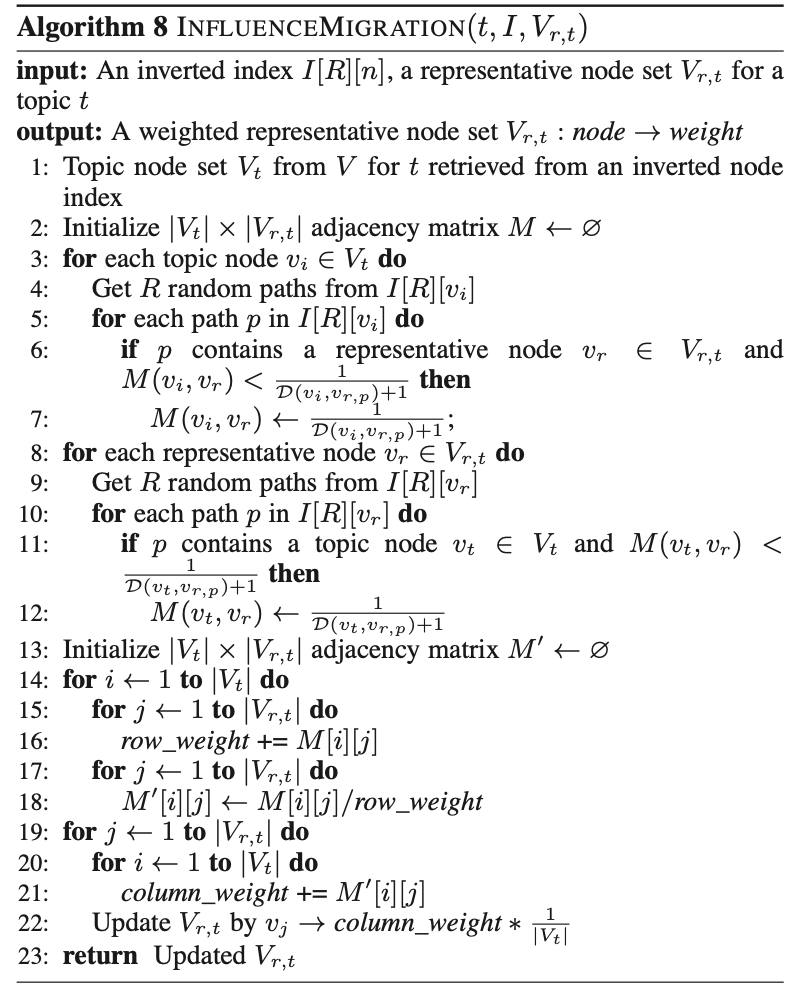
这一节主要任务是在上述的聚类算法的情况下，将话题节点的局部影响力有效地迁移到相应的话题相关代表节点上。影响力迁移的关键思想在于：给定一个主题节点，从该主题节点开始生成若干条随机行走路径。对于每条随机路径，要检查它是否包含任何代表节点。如果有，遇到的第一个代表节点就被视为吸收马尔可夫链的吸收状态。一旦进入吸收状态，就不能再离开。在一个主题节点的所有随机路径被遍历之后，根据所有随机选择的路径，确定一组与主题节点局部接近的代表节点。然后对每个代表节点重复这个过程，找到所有先遇到的和被吸收的主题节点。这里，处于被吸收状态的主题节点代表它将被路径的起始节点吸收。通过反复确定前向和后向吸收，发现每个主题节点最近的代表节点的可能性是最大的。然后，根据每个主题-代表节点对之间的接近概率分布，将每个主题节点的本地影响权重迁移到本地代表节点上。

为了计算话题-代表节点对的接近性概率分布，使用了一个关联矩阵M(i,j)。每个元素代表当前话题节点i和当前代表节点j之间的距离：[D(i, j)+1]-1。距离越短，话题节点和代表节点之间的密切程度就越高。因此，代表节点将从话题节点i吸收更多的局部影响。在处理完所有话题节点和代表节点的随机路径后，矩阵M归一化为



假设话题t有#Rep个代表节点和m个话题节点，即可计算矩阵每列平均值得到每个代表节点j的聚合影响力，其中每个主题节点都设定为统一的局部影响力1/m：





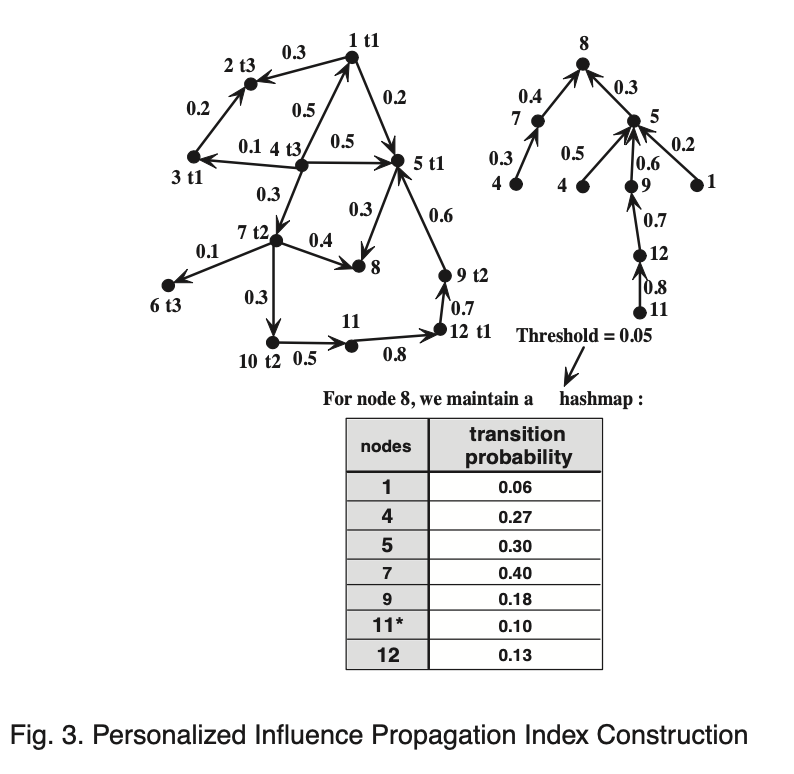
**5.** **主题的个性化影响**

在这一节中，我们提出了一种物化算法，通过评估代表节点的影响，有效地计算出每个用户的个性化top-k主题。要做到这一点，必须克服两个重要的问题：(1)每个节点都必须被物化；(2)需要一个高效的top-k算法，该算法是针对主题的。

**（1）个性化传播影响指数**

与寻找图中最短路径的一般问题不同，主题的个性化搜索只需要社交图中给定社交用户节点的 "附近 "节点。这里，"附近 "是指能够以高于固定阈值的过渡概率到达给定社会用户的节点。由于传播路径的过渡概率随着路径距离的增加而降低，在个性化搜索中只选择附近的节点。因此，图中所有节点对的最短路径是不必要的，只有每个起始位置的一小部分节点必须被物化。

物化的关键思想如下。对于G中的每个节点v，最初选择v作为基于后向的树的根，根据设定的阈值进行反向广度优先搜索（BFS）。



考虑图3中的例子，有12个节点，每条有向边上有一个过渡概率。对于每个节点，记录了当过渡概率高于θ时，在单一路径遍历中能够到达某个节点的附近节点。假设θ=0.05，起始节点为节点8。基于从节点8开始的反向广度优先搜索，将产生图3中第二个图的树。通过汇总每个节点的不同路径的影响，可以生成图3中的查找表。

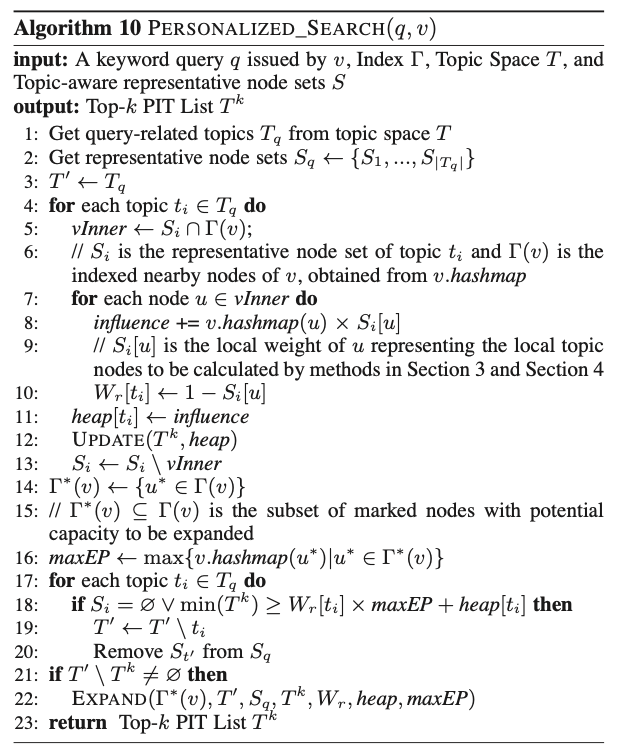
**（2）计算顶级个性化有影响力的主题**

给定一组与v发布的查询q相关的主题T，top-k PIT-Search算法的关键思想是首先通过探测v的物化节点索引和T的代表性节点集来选择top-k主题候选者。在v附近出现的所有代表性节点都完全处理完毕后，如果主题不能在基于总影响分数上限的前k主题候选人中，则可以从T中修剪一些主题。

修建规则：

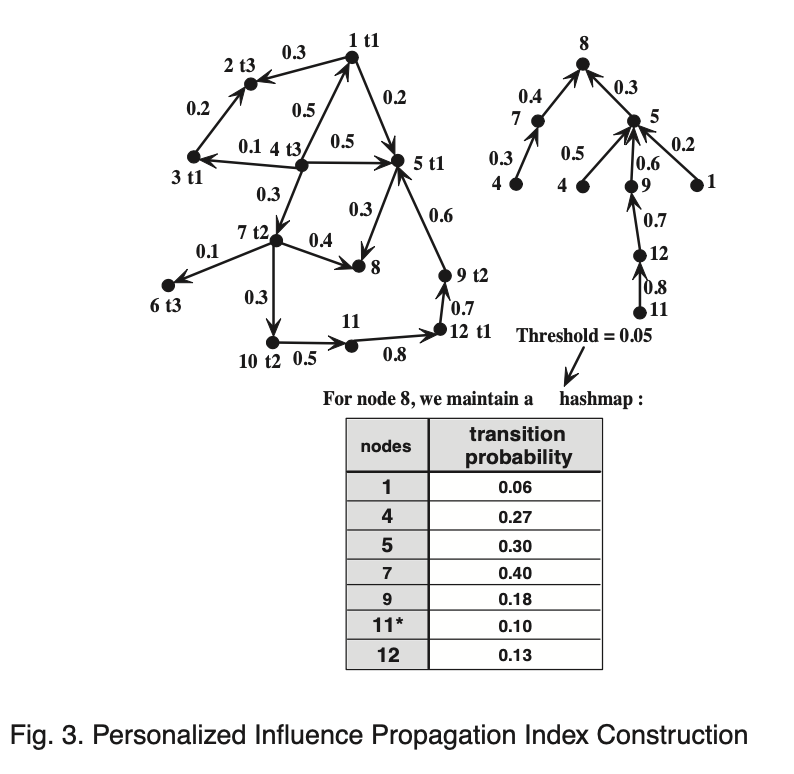
(1) Si主题中没有剩余的代表节点；

(2) 最小值min(Tk)大于或等于ti的影响力上限，其中ti的上限值是当前主题计算的影响力，以及剩余代表节点的最大可能影响力的总和。



以图3算例为例，假设节点8发出一个与三个话题t1、t2和t3有关的查询，并将k设置为1。Tq = {t1,t2,t3}和Sq = {S1 = {1,3,5,12}, S2 = {7,9,10}, S3 = {2,4,6}}被实例化。这里，Γ(8)={1,4,5,7,9,11, 12}。对于t1，vInner = S1 ∩Γ(8) = {1, 5, 12}。对于节点集vInner中的每个节点，计算对节点8的影响。如果假设S1中的每个节点都有0.25的本地权重，那么{1, 5, 12}相对于节点8的聚合影响力是8.hashhmap(1) × 0.25 + 8.hashhmap(5) × 0.25 + 8.hashhmap(12) × 0.25 = 0.055。（0.1225？）同样，t2相对于节点8的影响被计算为8.hashhmap(7) × 0.33 + 8.hashhmap(9) × 0.33 = 0.19。对于t3，对节点8的影响是8.hashhmap(4) × 0.33 = 0.09。

在这个阶段，中间结果是heap = {t1 = 0.055;t2 = 0.19;t3 = 0.09;}, Tk=1 = {t2}, S1 = {3}, S2 = {10}, S3 = {2,6}。根据中间结果，对不能成为top-1话题的话题进行修剪。由于S1和S3不是空的，它们的影响上界被计算为Wr[ti] × maxEP + heap[ti]。对于t1，上界是Wr[t1] × maxEP + heap[t1] = 0.25 × 0.1 + 0.055 = 0.08。因为t1的上限（0.08）小于t2的当前影响值，所以t1可以从进一步的计算中修剪掉。同样地，t3也可以被安全地剪除，因为其上界是0.157。因此，t2被返回作为节点8的前1个个性化的有影响力的主题。

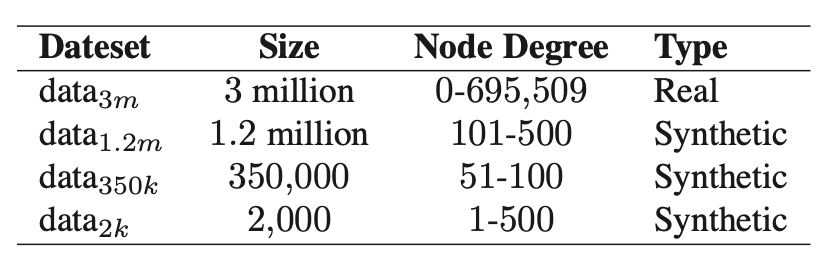


**6.** **实验**

**（1）数据集**

推特数据集：使用大型数据集（data3m，包括300万社交用户）来评估PIT-搜索的性能，其中包含2.84亿个 "关注 "关系，300万个用户资料和5000万条推文。

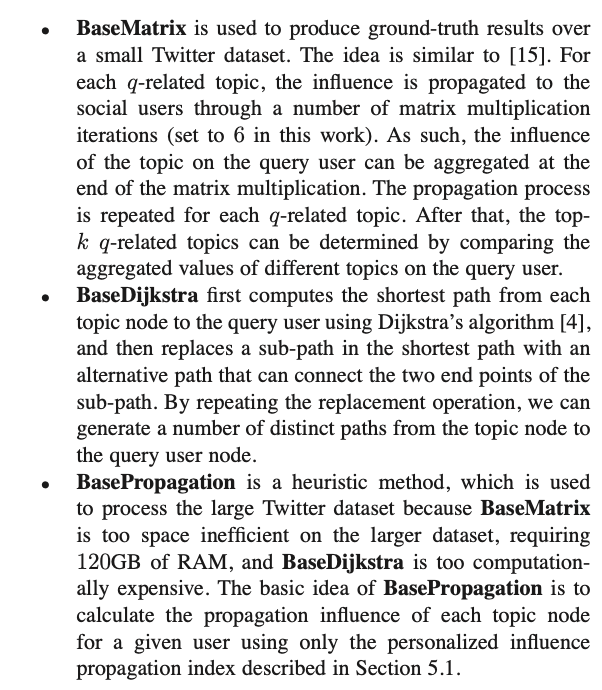
此外，从更大的数据集中生成合成数据集。使用类似的节点度数分布，从度数范围为51-100、101-500和500-1000的节点中分别产生三个合成数据集。这三个数据集分别表示为35万个节点的data350k，120万个节点的data1.2m，以及一个小的数据集（data2k），其中有2000个从原始数据集中用随机选择的社会用户。为了确保每个生成的数据集是一个连接的图，在断开连接的部分的接近节点之间添加了几条合成的边。



**主题生成。**给定一个Twitter用户，首先将其发布的信息视为一个文档，并将一个简单的LDA主题模型应用于该文档，生成一个术语包（通常为16个术语），作为该用户的主题种子。通过这样做，我们为每个Twitter用户准备了一套合理的主题种子。

**（2）对比方法**

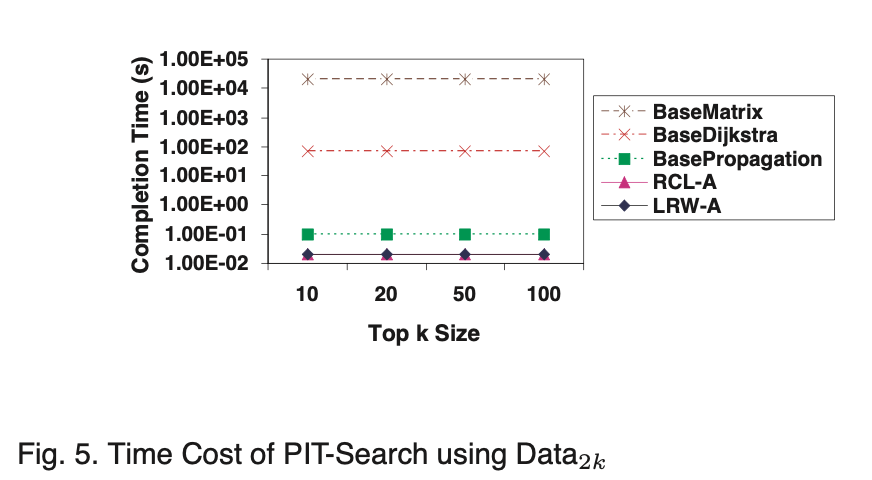
为了评估新的PIT搜索方法（RCL-A和LRW-A）的有效性和效率，使用三种不同的基线。

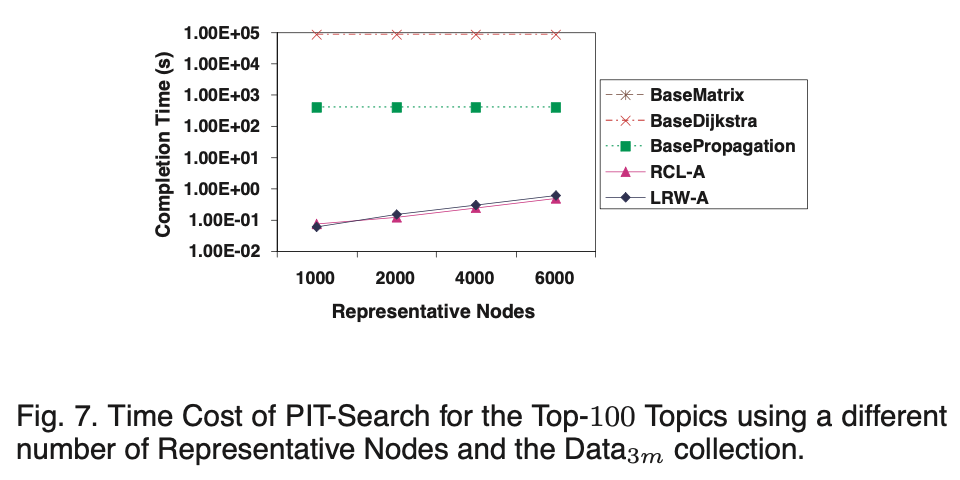


**（3）性能比较**

**PIT搜索效率**

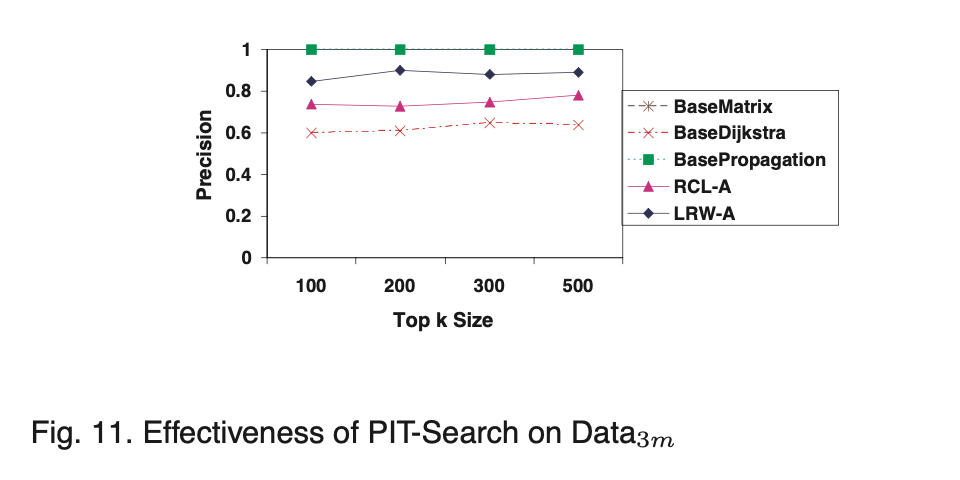
为了测试PIT搜索性能，选择了100个标签来表示用户的关键字查询。每个标签将在主题生成过程中为Twitter数据集生成500多个主题。然后，随机选择额外的49个用户，但保持100个采样关键字查询不变。所有运行的平均值用于公平评估PIT搜索性能。



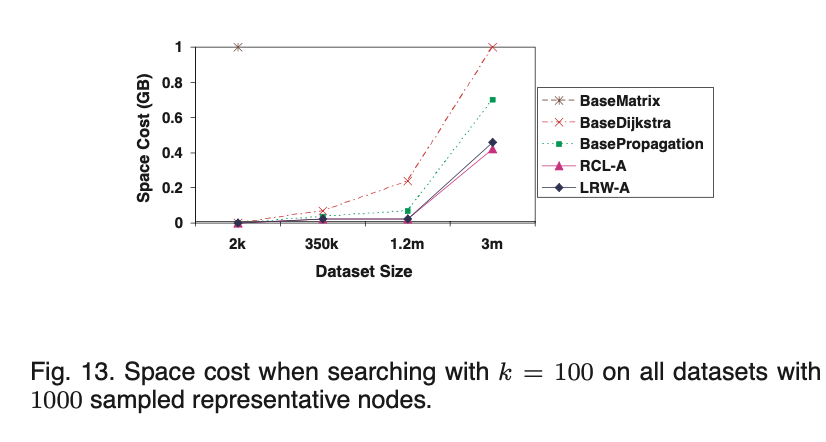


**PIT搜索有效性**

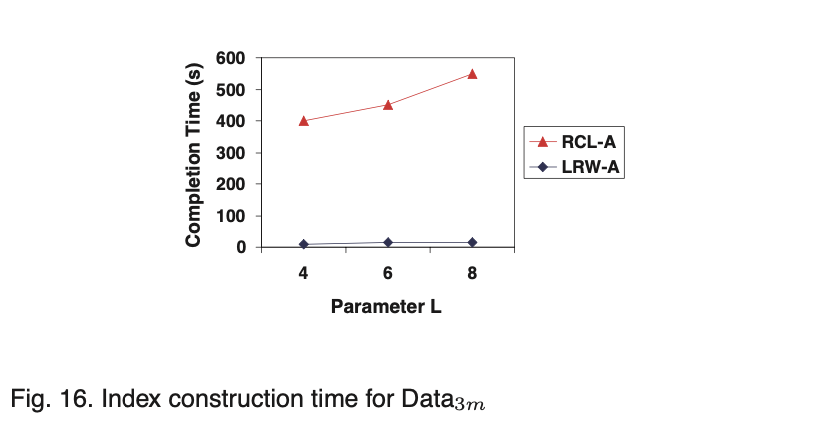
为了衡量PIT搜索近似方法的有效性，我们将BaseMatrix生成的结果视为小型Twitter数据集的基础真理，因为它详尽地支持停止所有路径。BaseMatrix根据主题的确切影响计算选择个性化有影响力的主题。



**PIT搜索空间成本**

****

**RCL-A与LRW-A对比**

****