**local experts finding Across Multiple social networks**

**1.introduction**

本地专家与一般主题专家的区别在于，本地专家的专业知识非常局限于地理位置。地方专家可以在很多应用中发挥重要作用，如解决地方信息查询、社会事件安排[5,18]、空间众包等。事实上，雅虎最近的一项调查显示，研究发现，人们更喜欢向当地的专家学习，因为他们熟悉社区，有第一手的经验。

之前的研究大多集中在单一来源的社交网络。然而，随着在线社交网络的快速发展，人们往往同时参与多个社交网络，享受更多的在线社交服务。从一个单一来源的社交网络中寻找本地专家，可能会遗漏一些重要的信息，导致被识别专家的不准确。

**挑战**：给定一个社交网络G，对本地专家进行的评分(本地专业知识)与本地社区的资讯有关。本地专业知识包含两个方面：地方权威，和话题知识。然而，由于存在以下挑战，传统的单一社交网络上的工作无法应用于本地专家在多个社交网络上的发现:

1. 如何在多个异构的社交网络上评估两种类型的价值，即地方权威和主题知识

2. 如何融合这两种类型的值，并做出权衡。

3. 由于社会信息在多个异构社会网络上的冗余性，如何有效地解决本地专家发现问题是一个迫切需要解决的问题。

**贡献点**：为了解决上述问题，本文提出了一个新的KTMSNs框架，他有两个步骤：首先，给定一个人在多个社交网络上，我们分别计算了当地权威和主题知识。作者将多个网络（m）的本地专家发现问题划分为m-1个子问题，每个子问题都是在两个异构社交网络上，针对每个子问题，我们提出了一种基于社会距离的知识衰减方法来评估两个异构社交网络中所有用户的局部权限和主题权限。

其次，作者提出了一个基于skyline的策略，结合这两个值来获得一个人的分数。通常，具有较高主题权限的人员可能位于远离查询位置的位置，因此本地权限可能相对较低。另一方面，拥有高地方权力的人可能拥有较低的局部权力。基于天际线的策略适合于解决这一权衡障碍。

最后，我们使用真实的数据集进行了广泛的实验评估，以了解所提索引的效率和所提算法的有效性。

**2. Preliminaries and Problem Formulation**

**2.1 preliminaries**

作者将在线社交网络建模为无向图G=(V,E)，每个节点v属于V都有一个地理坐标（x,y）

定义1：Binary Aligned Heterogeneous Social Networks：如果两个不同的社交网络共享一些共同的用户，那么这两个网络就称为对齐网络，我们将两个对齐的社交网络命名为二进制对齐的异构社交网络，可以表示为：，是两个网络中无向锚链接的集合。

定义2：锚链接：假设Vi和Vj是网络Gi和Gj的用户集合，是同一个用户在两个不同网络中的账户， 就是两个图中的无向锚链接。

定义3：锚用户：在两个网络中都有账户的用户。

定义4：全对齐：如果两个网络中有至少一个网络中的用户都是锚用户，那么就称这两个网络是全对齐的。

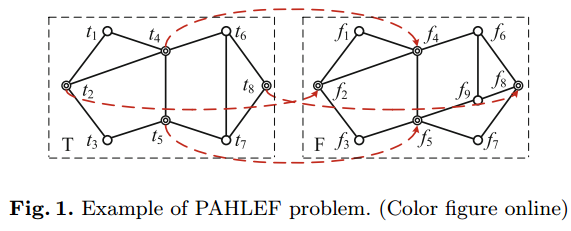
定义5：部分对其：在两个网络中都有用户为非锚用户。

**2.2 problem formulation**

定义6：Local Experts Finding over Partially Aligned Heterogeneous networks, PAHLEF：给定一个部分对齐的异构社交网络



并且给定一个查询，PAHLEF的目标是找到一个由k个候选人组成的集合，这些候选人在查询地点t(q)和地理位置l(q)中有很高的权威。

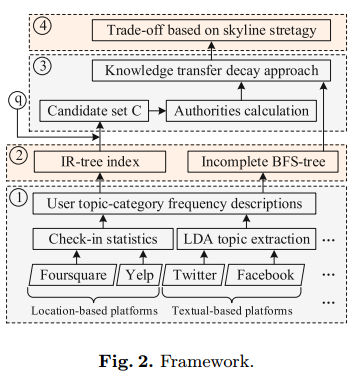


例子1：上述两个网络是部分对齐的，有两个圆环的是锚节点，红色的线是锚链接。假设给定一个查询是，目标就是在两个网络的用户集合中找到4个在music和地理左边l（q）中最在行的人。

**3. approach**

**3.1 overview**

我们详细介绍了我们提出的KTMSNs方法，其框架如图2所示。KTMSNs由数据准备阶段、社会拓扑感知反向索引构建、知识衰减转移和权衡讨论组成。我们将不同类型的社交网络数据转换为统一的指示。对于基于位置的平台(如Foursquare、Yelp等)，常见的用户行为是在不同的地点签到。场地有它们的主题类别。我们可以通过统计分析将用户行为描述为主题类别频率，我们利用LDA模型[3]和基于TF-IDF的方法[29]来提取主题信息。在这篇论文中，我们关注的是本地专家在多个异构社交网络上发现问题。基于这种直觉，我们将主题类别频率描述作为数据准备的一部分。在本文的其余部分，我们将不做详细说明。



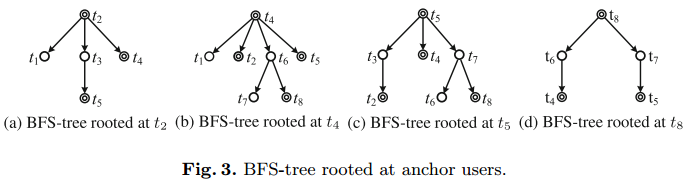
**3.2 Index contruction**

Social Distance Oracle：给定部分对齐的异构社交网络，我们为每个锚用户在其社交网络中构建一个BFS树。但是为每个锚用户都构建一个完全BFS树是多余的，作者使用下面的理论1来缩减树的规模。

理论1：如果这个分支中有另一个锚用户，则可以停止扎根于任何锚用户的BFS-tree的每个分支的扩展。

oracle的社会距离索引包含不完整的扎根于不同锚用户的bfs -tree。因此，我们可以有效地查询任意非锚用户u和任意锚用户v之间的社会距离。我们用dist(u, v)表示社会距离。我们首先遍历以锚用户v为根的BFS-tree Tv，如果非锚用户u位于此BFS-tree Tv中，则很容易得到社会距离。否则，我们得到一个锚用户集合A(Tv)，其中所有顶点都是出现在Tv中的锚用户，我们遍历扎根于属于A(Tv)的锚用户的bfs -树。一旦我们在任何BFS-tree上找到非锚用户u(假设这个BFS-tree的根是锚用户w)，我们就将u和v之间的社会距离更新为dist(v, w)和dist(w, u)的和。

例子2：以图1中部分对齐的异构社交网络为例。为了简化描述，我们只给出一个关于Twitter社交网络的简单例子。如图1所示，Twitter社交网络中有4个锚用户，我们构造了以每个锚用户为根的BFStree，如图3所示.



作者以t3和t8为例来展示如何利用社会距离oracle计算非锚用户和锚用户之间的社会距离。我们首先遍历以t8为根的BFS-tree，检查t3是否在该树中。t3不在T(t8)中，因此我们可以得到锚用户集合，接下来，我们遍历以A(T(t8)中的锚用户为根的BFS-tree，

遍历以t4为根的BFS-tree之后，锚用户设置A(T)更新为{t5, t2}。们更新了t3和t8之间的社会距离为与另一条社会路径t8→t4→t2→t3的社会距离大于3。最后，t3与t8的社会距离为3。

**Topic Category-Aware Inverted Index**：索引的另一个组成部分是主题分类感知的反向索引，对于每个社交网络平台中的Gi，作者构建了一个ir树[4,8,22]来索引与社交用户相关的空间信息和主题类别频率。

主题类别频率反映了用户的活动体验，反映了用户所掌握的知识。主题类别层次树可以从维基百科中提取，ir树本质上是一个用反向文件[29]扩展的R-tree[11]。在本文中，ir树中的每个叶节点都包含一个表单(u, u.λ,u.di), u指的是社交网络中的用户，u.λ是u的边界矩形, u.di是描述u的标识符。每个叶节点还包含一个指向反向文件的指针，该文件包含存储在节点中的用户的主题类别知识。

反向文件索引有两个主要组件(1)所有不同主题类别的词汇表;(2)每个主题类别c的发布列表，该列表是包含c经验的用户的键值对序列和相应的主题类别频率。

ir树中的每个非叶节点R都包含一些表单条目，cp是R的子节点的地址，cp.λ是子节点条目中所有矩形的最小边界矩形(MBR)，cp.di是伪文本描述的标识符，伪文本描述是子节点条目中的所有文本描述的联合。此外，非叶节点的反向文件是其子树的总结。具体来说，我们将与同一主题类别相关的最大知识得分记录在一个MBR中。

以图1中的Twitter社交网络为例，我们假设用户的空间分布如图4(a)所示。图4(b)显示了用户的描述，其中第一列是用户列表，第二列包括主题类别和频率。

图5显示了相应的ir树。表1说明了与节点关联的反向文件的内容。值得注意的是，我们为每个社交平台构建了一个IR-tree索引，而不是构建一个整体的IR-tree索引。直观上是评估所有在线社交网络中每个社交用户的本地权威和主题权威。对锚用户建立一个综合的IR-tree指数，用于两个权威机构的计算是有益的。但是，它不适合非锚用户。因此，我们提出了一种基于社会距离的知识衰减方法来估计多个异构社交网络中所有用户的两个权威。一个整体的IR-tree指数对于一个社交网络到另一个社交平台的知识转移是不重要的。我们将在3.3节中详细阐述知识衰减法。

**3.3 Knowledge Decay Across Social Networks**

作者将基于多个社交网络(假设有m个网络)的本地专家发现问题分开为m-1个子问题，在不失一般性的前提下，我们分析了关于本地专家在两个部分一致的异构社交网络中发现问题的建议方法的细节。首先介绍了如何计算两个权威值，然后详细阐述了两个异构社交网络之间的权威转移过程。

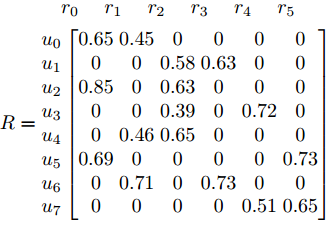
**Local authority**：当一个本地专家发现查询q= <t(q), l(q), k>被提出来，直观上，靠近l(q)位置的“追随者”越多，候选人的地方权威就越高，这个计算可以按如下公式来计算：



代表用户u的local authority，vi是u的一个追随者代表vi和l(q)的空间距离，表示查询位置l(q)周围的半径。我们利用r树结构来加速local authority的进度。如图4(a)所示，星形为查询位置，灰色虚线圆圈为半径为r(l(q)的区域。

**Topical authority：**了评估主题权限，我们首先利用主题类别感知的反向索引，通过将查询主题t(q)映射到主题类别层次树来生成候选集C。如果树中的映射节点是一个非叶节点，我们将该非叶节点的子树作为查询主题扩展。例如，是主题类别层次树中的非叶节点，我们将查询主题扩展到，其中每个元素都是子树的一个节点。我们从反向索引的词汇表中检索这些元素，每个拥有t (q)中任何元素的不可消失体验的用户都包含在候选集C中，然后分别计算这两个社交网络中锚用户的权限。因此，我们将锚用户放入候选集合C中，因为所有锚用户都需要计算两个权威机构的得分。用户的话题权威与活动体验E具有相互强化的关系。在主题区域中拥有高权限的用户将在主题区域中处于活动状态。专题领域的活动将加强相应的专题权威。更具体地说，可以通过集成用户生成的活动体验来计算用户的主题权限。与主题区域相关的活动经验可以通过统计主题类别频率来计算。

给定一个活动记录，我们可以在用户和主题类别之间建立一个相邻的矩阵R。在这个矩阵中，item rij表示主题类别频率，0 ≤i <|U|， 0 ≤ j < |T|。例如，图4(b)所示的矩阵可以表示为：（假设主题类别集表示为）



因此，用户局部权威的相互强化关系，激活经历可以用下面的式子表示：



**Authorities Transfer**

在单个社交网络上有效计算出这两种权限之后，我们准备对Bi-AHSNs上的每个候选c∈c对这两种权限进行评估。如果c是锚用户，我们将分别对这两种类型的权限进行求和。相反，如果c是一个非锚定用户，我们提出了一个基于社会距离的指数衰减模型来估计这两种权威。指数衰减模型的主要思想是知识可以通过锚链从一个社交平台转移到另一个社交平台。因此，锚用户是知识转移的中介。如图6所示，有两个社交平台A和B。为了评估非锚候选用户(假设c)在两个部分对齐的异构社交网络上的权限，我们首先在c所在的平台上找到c与所有锚用户之间最短的社交距离，可以通过子节3.2所示的BFS-tree索引得到。假设两个社交平台之间的锚用户集表示为AAB(缩写为A).因此，得分估计如下:



**3.4 Trade-Off discussion**

尽管我们可以在多个部分对齐的异构网络上评估这两种类型的权限，但仍然存在一个障碍。为了满足查询结果的多样性，我们需要返回k个候选项的集合，这些候选项具有与查询主题t(q)和位置l(q)相关的最高本地专业知识。如何在地方权威与地方权威之间做出取舍是一个挑战。在本文中，我们将两个权威机构视为本地专家的两个维度，而众所周知的skyline[2,17]策略适合处理两个维度的权衡问题。

skyline策略对于涉及多准则决策的几个应用程序非常重要。给定一组对象该策略返回k个对象作为top-k查询的答案，这样返回结果中的任何对象oi都不会被另一个对象oj控制。在本文中，多准则是局部权威和局部权威。domination的定义如下。

**定义7：domination**：给定一个用户集合我们把这两类权威看作是两个正交维。二维空间中的用户界面可以控制位于左下子空间中的用户。我们把这样一个子空间表示为ui的支配区域。

用户界面的控制数量是用户界面控制区域内的用户数量。因此，我们按照控制数的降序对候选集合C中的用户进行排序。最后，返回前k个用户作为本地专家查找问题的答案。

**4. 实验**

**4.1 实验设定**

**数据集**：我们在实验中使用了两个真实的数据集

Foursquare：通过Foursquare API抓取Foursquare数据集。2014年11月至2016年1月，拥有76,503名用户和1,531,357条社交关系。对于每个用户，我们收集了这段时间的check-in和地理位置。每次签到都是一对用户和地点。我们总共爬了299,995个位于新加坡的网站和969,549个签到站点。每个场馆都有自己的品类类型描述，如博物馆、中餐厅等

Twitter：我们通过Twitter API3收集了2014年11月至2016年1月的Twitter数据集，其中包含160338名用户和2405628条社交关系。对于每个用户，我们收集了一组最近(≤1000)的用户生成tweets和地理位置。

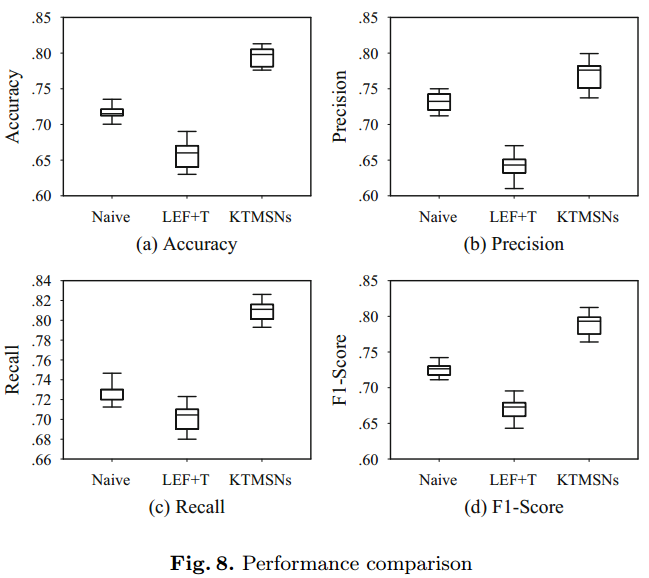
在这两个数据集中共有15281个普通用户，即这些用户同时拥有Foursquare账号和Twitter账号，并且对应的账号对已经被一些强有力的证据所匹配，比如他们是通过相同的电子邮件注册的。同时，利用数据准备阶段描述的方法发现用户的主题类别频率。图4(b)给出了一个简单的例子。

查询：作者的查询模型定义为,我们随机选择50个查询关键字和查询地址，将k = 10设为默认值，对算法性能进行评估。这些查询组合由20个一般主题查询(如“技术”)和30个特定主题查询(如“川菜美食家”)组成。我们将查询位置集分配为{the West of Singapore, the East of Singapore, the South of Singapore, the North of Singapore, the Whole Singapore}这意味着我们在每个通用查询主题下有四个通用查询主题和六个特定的查询主题。

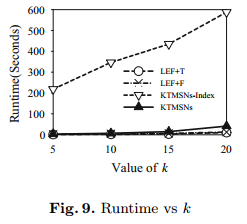
**对比算法**：Naïve，LEF+T，KTMSNs

**4.2 实验结果和分析**

性能比较：



效率评价：



k的影响：

