**Abstract**：具体来说，我们专注于预测用户的家庭位置、推特位置和提到的位置。我们首先定义这三项任务，并审查评估指标。通过总结推特网络、推特内容和推特上下文作为潜在的输入，我们从结构上强调了问题是如何依赖这些输入的。此外，我们还简要回顾了两个相关问题，即语义位置预测和兴趣点推荐。最后，对调查进行总结，并列出未来的研究方向。

**1.Introduction：**

在线社交网络急速发展，在这些平台上，用户可以与其他有相似兴趣的人建立在线友谊。用户也可以通过短信、照片、视频或签到的形式与在线朋友分享他们的日常生活。

**为什么选择Twitter**：1.推特上的友谊不一定是相互的。例如，用户可以“追随”名人，而不要求他们追随。2. 另一方面，推特(又名推特或微博)上的文字帖子限制在140个字符以内。

虚拟网络与现实网络之间的媒介：位置。用户有长期居住地址，实时地址。本文关注三种地址：用户家庭地址，tweet地址，提到的地址。

Twitter信息不完整，首先要补全信息。直觉上，推特相关位置的识别和消歧也应该在很大程度上依赖于推特文本。生活在某些城市的用户可能会讨论当地的地标、建筑和事件，可能会使用方言或俚语。

**文献目标**：在这次调查中，我们的目标是在推特上完成位置预测问题的总体描述。

**2.问题综述**

**2.1 Twitter综述**：

三种数据：1）用户发布的短而嘈杂的推文，2 )用户间建立的庞大的推特网络，以及3 )用户和推文的丰富的上下文信息。这些信息作为输入，能够研究一些地理定位问题

推文内容，可以自己写也可以转发别人的，使用@来提醒别人看

推特网络：ui follows uj。Ui是关注者 uj是被关注者。只要一方关注另一方，我们就称之为朋友，二者互关则称之为mutual friend

据观察，现实生活中的朋友往往会在网上频繁交往

在介绍关于暗示现实生活中朋友关系的线索的研究时，我们考虑以统一的方式跟踪和提及推特用户之间的行为，并将由此产生的网络称为推特网络。

推文上下文：不光有内容，还有时间标签，可以选择地理标签，还有用户信息。

**2.2地理位置预测**

预测三种地点：家庭住址，推文地址，提到的地址。

**家庭住址预测**：

家庭所在地指的是推特用户的长期居住地址。

家庭住址的三种粒度：行政区域：国家，州府，城市；地理网格：地球被划分为好多格子；地理坐标：经纬度。

一般来说，用户自己填的个人信息上有地址。但是隐私问题，很多人是空的。有些研究用推文标签来聚合出住址信息。可能的聚合方法为：1.最常见城市 2.第一个有效地理标志 3.地理标签的几何中值。

**推文地址预测**：位置通常基于tweet的地理标签。兴趣点或者坐标被用来表示推文地址

**提到的地址预测**：到的位置识别，即在提及位置名称的推文中提取文本片段。提到的位置歧义消除，即通过将这些片段解析为位置数据库中的条目来识别这些片段所指的位置。使用行政区域或者兴趣点

**2.3推特对位置预测问题的输入**

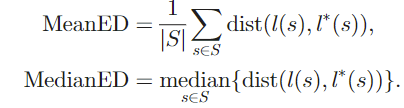
推特上的所有三种类型的信息，即内容、网络和上下文，通常被用来解决三个位置预测问题，即家庭位置、推特位置和提到的位置的预测。

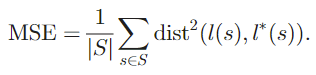
**2.4评估指标**

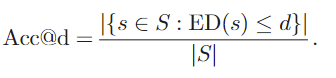
基于距离：位置由地理坐标表示

ED：误差距离定义为地面真实坐标和预测坐标之间的欧几里德距离:

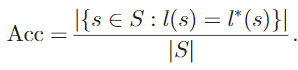




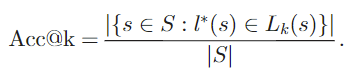


 设置一个阈值，将误差小于这个阈值的加入进来

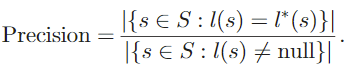
**基于令牌：**将位置视为离散符号，国家城市区域POI

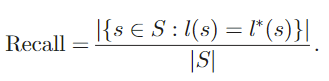
准确度：

在某些情况下，系统可以给出预测位置的排名列表，而不是一个。如果正确位置在前k个结果内，就认为是正确的



最后，我们注意到，在某些情况下，地理定位系统可能无法做出预测。例如，在家庭和推特位置预测中，如果信息不足，一些系统无法分配位置。在这种情况下，精确度、召回率和F1被用作度量标准。







F1是二者的调和平均值

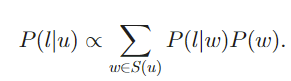
**3家庭位置预测**

在大多数研究中，家的位置是在城市一级预测的，有时是在州或国家一级。

**3.1基于推文内容的推理**

用户的家的位置可以通过推特内容中的某些词语随意透露。基于内容的家庭位置预测的潜在挑战是通过那些指示性单词将用户精确地链接到位置。

**1.word-centric** 单词中心法旨在识别和利用这些单词来预测用户的家庭位置（休斯顿火箭，俚语）**。**因此，应该只涉及本地单词，即显示强局部性的单词。此外，在作出预测之前，应该从数据中学习本地单词所隐含的位置信息或它们的空间单词用法。

**识别本地词汇**，至于推特，通常情况下，像“市区”和“OMG”这样与位置无关的词比像“你好”和“费城人”这样的词出现得更频繁。如果不加区别地考虑，它们将导致定位预测结果随机。我们通常采用消除位置无关词的方法，即识别和保留本地词。由于在大多数应用中，本地词不像停止词那样是可枚举的，所以大量的研究工作都花在识别本地词上，不管是无监督的还是有监督的。**无监督**的局部单词识别方法旨在对数据进行直接计算并指示单词位置的统计度量。Laere等人提出了两种本地选词方法。一种是利用核密度预测，另一种是利用里普利K统计。受到IDF的启发，Ren et al. [36] and Han et al.提出了ILF和ICF，他们的假设是本地单词应该分布在更少的位置，并且具有更大的ILF和ICF值。**监督的**方法。Cheng等人将本地词识别问题视为一个分类问题，首先，他们用Backstrom等人的空间变异模型拟合每个单词的地理分布。模型拟合后，参数用作词特征。其次，他们在字典中手工标注19，178个单词为本地或非本地。最后，他们训练了一个分类模型，并将其应用于tweet数据集中的所有其他单词。  
 **空间词应用**：识别本地单词后，下一个问题是如何使用它们来预测用户的家的位置。程等人介绍了一个有代表性的概率模型

Li等人[ 40 ]观察到，发布场馆名称的概率在某些时候是基于位置的，而在其他时候也是随机的。因此，他们将其作为两个级别的评估。采用伯努利分布来估计位置名称是随机发布还是基于位置发布，然后使用多项式分布来估计从每个位置发布地点名称的概率。

Cheng等人 第一种方法，拉普拉斯平滑法(或加一平滑法)，将所有位置的字w的计数增加一，然后归一化以产生分布。随后状态级平滑和基于网格的邻域平滑。

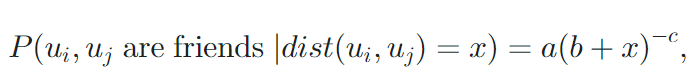
**2. location-centric methods** 一些研究采用基于分类的方法进行家庭位置预测。他们将用户关于本地单词的统计数据视为特征，并将所有候选位置视为分类标签。Hecht等人[ 8 ]选择CALGARI得分最高的前10，000个单词作为本地单词。然后，用户被表示为10，000维术语频率向量，并被馈送到多项式朴素贝叶斯分类器中，用于训练和家庭位置预测。同样，Cha等人。[ 44 ]使用稀疏编码和字典学习技术进行单词特征选择。在他们的扩展工作[ 45 ]中，他们还建议从训练数据中识别和移除旅行人员，以提高归属位置分类器的性能。如果一个人的两条推文是从距离超过100英里的地方发出的，那么这个人就被认为是在旅行。除了传统的方法外，最近的一些研究还探索了深度学习模型来解决家庭位置预测问题。通过扩展他们以前的工作[ 49 ]，三浦等人。[ 50 ]提出了一个更复杂的模型。他们按时间顺序排列用户的消息，并应用顺序模型RNN对内容进行编码。

**3.2 基于推特网络的推测**

用户的社交关系也会揭示他们的住址，因为朋友间大都住的较近。

**3.2.1 基于友谊的方法**

一个人的家庭住址肯定是跟它的朋友的住址是有关系的。Ren认为一个人的朋友中在同一个地址住的比例越高则这个人也在这个地址住的几率就越大。上述方法的一个缺点是它们不使用用户朋友家的坐标。位置被视为一组离散的对象，而它们之间的距离被忽略

Backstrom分析了大量具有已知家庭位置和朋友关系的脸谱用户。他们试图用下面的曲线来拟合两个用户成为朋友的概率为

换句话说，友谊的概率与家庭距离成反比(截距为b )。

以上方法换句话说，他们含蓄地假设在线社交网络上观察到的友谊意味着真正的离线友谊，从而拉近了家庭距离。但是这在现实中时不正确的

孔等人。发现一对朋友有83 %的机会住在10公里以内，如果他们的普通朋友分别占他们朋友的一半以上。如果普通朋友的比例限制在10 %，这种几率会降低到2.4 %。这意味着推特上丰富的间接好友关系可能更好地表明两个用户之间的离线友谊，从而表明他们的家在附近。科斯涅茨等人也观察到了这一点。[ 58 ]，如果两个用户a和b与许多第三用户有关系，则a和b可能有关系。三浦等人[ 50 ]将用户友谊信息编码成神经网络模型。

**3.2.2 基于社会亲密度的方法**

在[ 57 ]和[ 59 ]的调查表明，社交亲密度，或者两个用户在现实生活中彼此有多熟悉，是家庭亲密度的更好指标。在推特网络中，提及是用户互动的另一种形式。当用户相互提及或相互交谈时，这两个用户被认为关系更密切或有着相似的兴趣。

McGee等人[ 23 ]对104，214名家庭位于美国境内的推特用户进行了分析。他们发现，除了通过关注建立的相互友谊之外，用户相互提及和积极聊天的行为也表明他们的家很近。在随后的工作[ 24 ]，麦基等人。通过检查更大的数据集证实了类似的观察结果。他们还进行了更多的观察: 1 )如果被关注的用户帐户是受保护的帐户7 (通常是普通人)，则两个用户在地理上很接近；和2 )当地报纸账户接近他们的关注者们。

康普顿等人。[ 25 ]也利用用户之间的提及。他们建立一个用户提及图，并优化未知的家庭位置，使相互提及的用户尽可能靠近。

Kwak等人分析了一个大型推特数据集。[ 60 ]发现拥有不到2000个共同朋友的用户(因此不太可能有很大的影响力)更有可能在地理上接近他们中的大多数。在本节前面描述的麦基的著作《[》中，他们还发现，拥有许多朋友和追随者的用户朋友往往离用户更远