**Abstract**：当前的研究都是基于粗粒度的，本文提出了一种基于用户签到模式的综合方法来预测用户在LBSNs中任何细粒度时间的未来签到位置。首先，分析了用户的签到模式，包括时间周期、全球受欢迎程度和个人偏好。其次，提取与用户签到模式相关的多个特征，并探索每个特征的预测能力。第三，将一组特征分别组合成监督评分模型和分类模型，用于预测用户在未来细粒度时间的签到位置。最后，在三个真实Foursquare数据集上进行了大量实验，验证了所提方法的有效性。

**1.introduction:**

移动互联网的普及加速了基于位置的社交网络的普及，其中一些典型的平台如Foursquare平台已经成为互联网用户不可或缺的。LBSN的特色是位置信息，用户可以通过这些信息与朋友分享签到信息，尤其是当他们访问有趣的地方时。LBSNs中数百万用户生成的大量登记数据可用于探索用户登记行为的内在模式，并基于他们的历史登记记录进一步预测他们未来的登记位置。事实证明，他的应用场景对于交通规划，产品推荐，灾难预测很有价值

许多早期的研究大多是基于全球定位系统的推测数据或者用户电话记录数据进行的。用GPS缺乏语义信息（比如用户在这个地方是参加什么活动）。至于从用户呼叫记录中获得的位置数据，它只能提供粗略地理信息。LBSNs (如Foursquare )中的登记数据更加精确，登记时间、协调以及社会关系也更加精确，此外，LBSNs中的位置经常被标注上类似类别的语义信息，便于研究者对用户签到模式进行多角度分析。因此，在LBSNs中开发基于大量签到数据的位置预测算法，保证了解决这一问题的通用性。

最近许多基于LBSNs中登记数据的研究(参见第2节)都是关于预测用户将要签到的下一个位置。他们不能预测长期以后的用户签到位置。对于一些提前很长时间预测签到位置的研究，他们往往单独关注个别特征，而忽略了各种因素对用户签到行为的综合影响。此外，现有的研究主要针对位置类别进行粗粒度的位置预测。显然，在日益激烈的基于位置的服务竞争环境中，这种位置预测方法的实用性被缩小了。

我们的目标是基于对时间周期性、全球受欢迎程度和个人偏好的综合考虑，预测用户未来在任何细粒度时间点的签到位置。本质上，我们的目标是预测用户在他们以前访问过的地方的未来入住概率。为了完成这一任务，我们首先从时间周期、全球流行度以及个人偏好等方面分析了用户签到模式。然后，我们将所有这些因素分别组合成一个监督评分模型和一个分类模型，从两个不同的角度来解决这个问题。

**2.related work**

**2.1 预测下一次签到地址**

现有的研究主要是关于预测用户下一个签到的位置。Lian等人( 2014 )彻底分析用户入住的属性以及用户的人口统计数据。位置可预测性和这两个方面之间的相关性表明，可以很好地估计用户的下一个登记地点。Gao等人( 2012 )提出一个全面的语言模型来描述下一个签到位置的生成机制，Noulas等人( 2012 )研究单个用户的登记模式预测下一个登记位置的能力，并进一步将一组特征集成在一起以提高预测性能。为了提高LBSNs中下一个位置预测的性能，Likhyani等人提出了一种新的定位方法。( 2015年)通过将稀疏的位置数据与地图信息相关联来探索粗粒度的场馆类别，然后利用该辅助信息对下次访问进行更准确的位置预测。

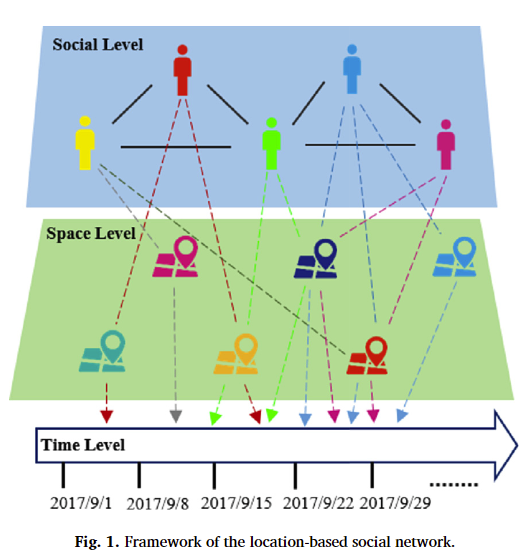
**2.2预测未来的签到地址**

很早以前就有一些关于位置预测的研究。庞和张( 2015 )证明了用户的移动性受到社会影响的限制，这种影响通过他所在社区的一小部分反映出来。考虑到用户的社区信息，他们提出的方法能够回答他是否会在给定时间在某个地方签到的问题。与我们在本文中的工作不同，它们忽略了时间周期的影响，而时间周期在决定用户未来的签到时也可以发挥重要作用。Assam和Seidl (2014 )通过小波量化了社会影响对用户签到的影响，以便对具有一个以上影响者的场景进行适当的审查。同样，没有考虑时间周期。

与上述研究相比，我们的研究具有对用户签到行为的时间周期性、全球流行度和个人偏好进行综合分析的特点，并且我们提出的方法能够提供精确到小时的位置预测，在基于位置的服务的激烈竞争中具有广阔的前景。

**3.问题定义：**

在我们的研究背景下，基于位置的社会网络由三个层次组成，即社会层次、空间层次和时间层次。社交层次包含用户与朋友的社交关系，空间层次包含所有历史登记发生的位置，时间层次分别指示何时进行登记。



一般的，一个LBSN可以由一个四元组G<U,V,E,C>组成。其中1 )U代表节点集，即网络中的用户；2 )V代表位置；3）E代表边的集合；4）C代表历史签到集合。如果一个用户u有n个签到记录，表示为

问题定义1：给定精确到小时的时间t，我们对所有Li（用户曾访问的地址）位置进行排名，用户ui将在时间t访问的准确位置被排列在排名列表中的最高可能位置。

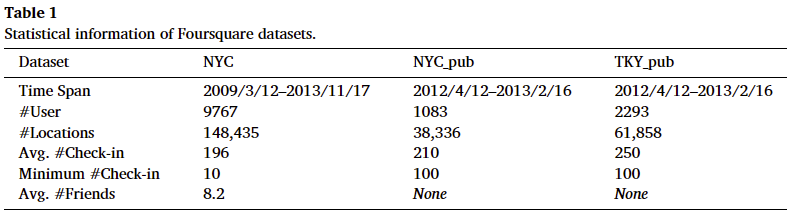
问题定义2：给定时间t和具体位置v，判断ui是否会在时间t在v签到。

应当注意，定义的两个版本对应于细粒度位置预测问题的不同方面。前一个定义侧重于对所有可能的候选位置进行排名，因此可以被视为一个排名问题。然而，后一个定义强调决定是否访问给定的位置，因此可以被视为分类问题。

虽然这两个问题可以独立处理，但我们发现使用评分模型和分类模型的组合可以提高预测性能。

**4.数据集**

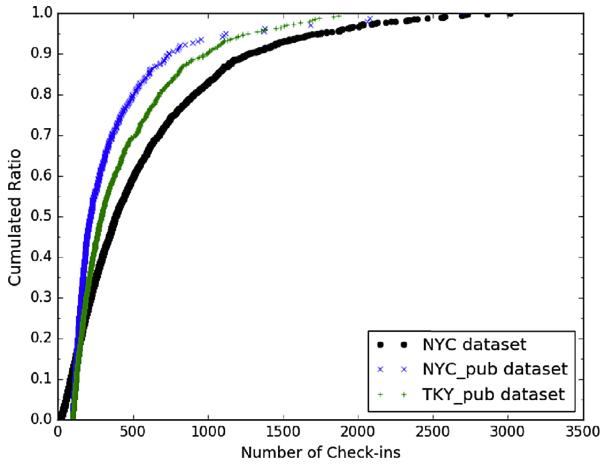
Foursquare数据集，他提供了充足的用户-位置信息和语义信息。Foursquare中的位置用一个层次分类系统进行分类，该系统涵盖娱乐、食物、旅行等方面。有了位置类别，我们可以很容易地获得可以用来推断用户兴趣的签到的语义信息。在这项工作中，我们使用了三个真实世界的Foursquare数据集，其中第一个NYC是我们自己爬取的，而另外两个数据集NYC\_pub, TKY-pub是yang总结的



至于NYC数据集，我们使用Foursquare的API和推特API对用户网络(所有用户都来自纽约市)和用户历史签到进行了爬取。数据集由三部分组成:社会关系、签到记录和位置类别。根据(叶等人。，2011年；Zhang等人，2012 )，签到次数和访问次数少于10次的用户将被删除

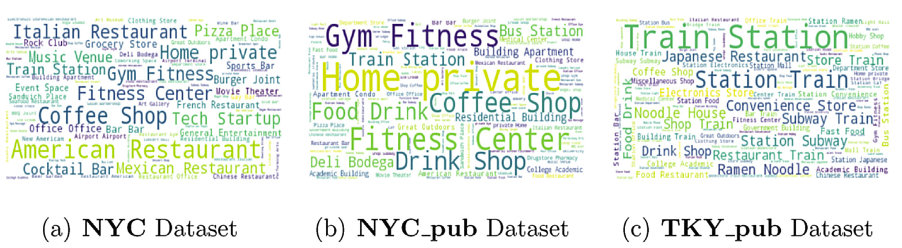
NYC\_pub和TKY\_pub分别指在纽约和东京的签到记录。尽管这两个数据集中的用户数量较少，但用户签到的最小数量和平均数量都高于NYC。社会友谊最初是在(杨等人)中缺失的。因此，我们不能使用这两个数据集来分析社会特征的预测能力。

下图展示了三个数据集的签到数量的累积分布



有1000个签到信息以上的用户很少。

我们还绘制了三个数据集上登记位置类别的标签云，如图3所示。请注意，位置类别通常嵌入有语义信息，可以将其视为用户个人偏好的表示。

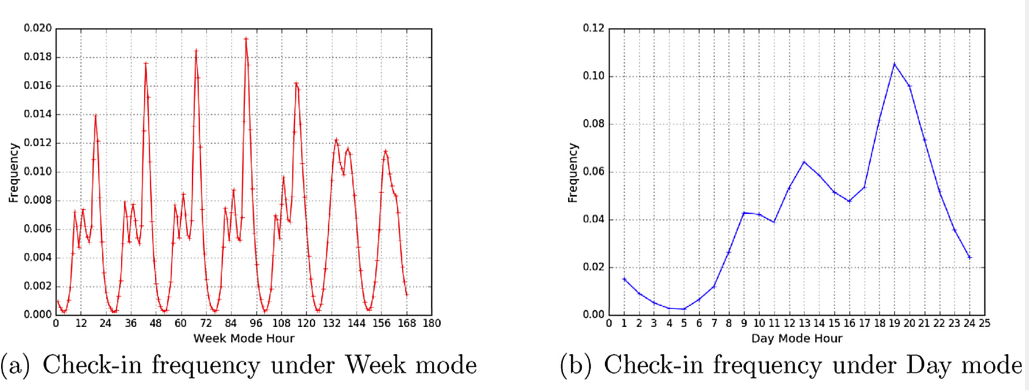


**5.预测特征分析**

在这一部分中，我们首先对涉及时间周期、全球流行度和个人偏好的用户签到模式进行了深入的分析。然后，我们探索单个特征对未来签到位置的预测能力。

**5.1用户签到行为分析**

人类行为会有很强的周期性，签到行为有可能在一天或者一周内循环。引入时间窗口(包括日模式和周模式)的概念来模拟用户签到模式的时间周期。具体来说，日模式是将时间映射到一天中的24小时，形成24个时间窗口，而周模式将一周映射到168小时。为了根据时间周期来分析用户签到模式，我们绘制了两种时间模式下的签到频率分布。



在周模式下，签到频率在工作日和周末之间变化。这主要是因为人们倾向于在不同的日子参加不同的活动。例如，在工作日，人们可能要在办公室、餐馆或咖啡馆打卡，然而，他们更可能在周末在家休息，这导致较低的签到频率。在日间模式下，用户的入住主要发生在白天，从上午9 : 00到晚上21 : 00，在晚上18 : 00左右达到高峰。

我们从位置类别的角度进一步分析用户签到构成。发现少数类别如食物或者购物往往占到了很大的比例，像一个很大的类别比如居住或者时间却占到了很小的比例。基于这一观察，我们推测四方用户倾向于分享与食物和购物相关的位置，这通常发生在晚上。这在一定程度上解释了为什么在白天模式下，入住频率在晚上达到峰值。与位置类别相关的另一个发现是，用户对不同类别有明显不同的偏好。例如，“学院和大学”场馆的报到时间是整周，但与工作日相比，周末的报到时间大幅减少。然而，对于“夜生活场所”来说，入住频率显示出完全不同的模式。用户倾向于在周末去这类地方。

Gao认为对于每一个用户都一个所谓的签到中心，当这人离这个签到中心越近，他就越可能在这签到。在本文中，我们采用了( Gao等人)中提出的方法。，2012 )来发现用户登记中心。我们绘制了NYC数据集中所有用户在双对数坐标系中的位移分布，如图5所示。这一趋势与现有的研究是一致的。

社交关系也是预测社交网站用户未来入住位置的一个重要因素

**5.2 特征抽取**

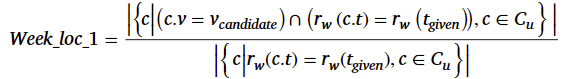
我们总共抽取12个特征

**5.2.1时间周期**



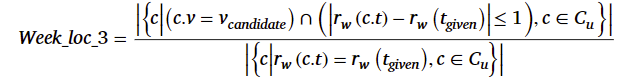
**特征week\_loc\_1:**



Cu是用户u的签到集合。week\_loc\_1相当于用户在周模式下候选位置的历史签到频率。、

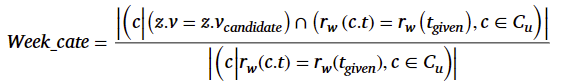
**特征day\_loc\_1**：跟上面的一样

**特征week\_loc\_3：**由于用户在同一地点的签到时间并不总是固定的，我们将给定的时间间隔扩大到3小时。



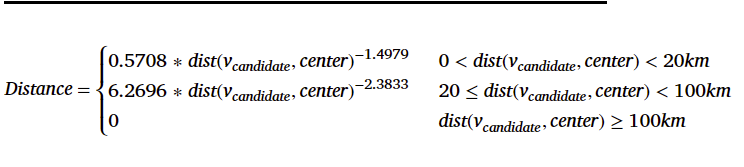
**特征day\_loc\_3：跟上面的一样**

**特征 Week\_cate:** z.v表示v类所属的位置。



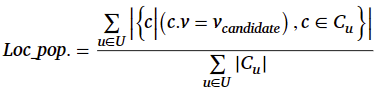
**5.2.2全局流行性**

基于用户经常在靠近其签到中心的位置签到的事实，我们定义个一个全局特征：候选位置的**距离**

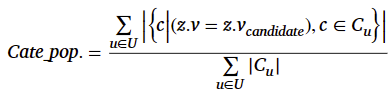


随后定义了另外两个全局特征：

**Loc\_pop:度量了候选地点的全局受欢迎程度**

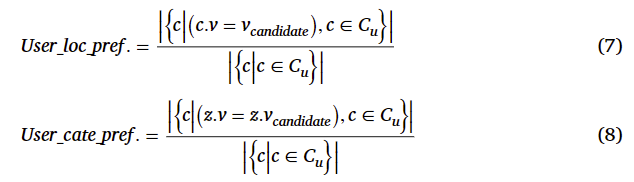


**Cate\_pop:**



**5.2.3个人偏好**

我们确认用户对特定位置和特定类别有自己的偏好，



最后，我们将朋友对用户的影响视为一种社会特征，通过它可以同时考虑用户之间的签到相似性和朋友的历史签到记录。



**5.3单个特征的预测能力**

给定用户的历史签到记录和精确到小时的特定时间，我们有一个位置集合L来排名。对于每个单独的特征，当给定特定时间时，可以计算每个候选位置的数值，该数值随后用于产生该位置的排名。我们用rankI(k)表示一个候选位置k在位置集合中的顺序。

为了衡量未来签到地点的排名，目前有两种主要方法：Average Percentail Rank和Accuracy@N

**Average Percentail Rank：**

可以看出在排名中越高的，就会有更高的PR值。

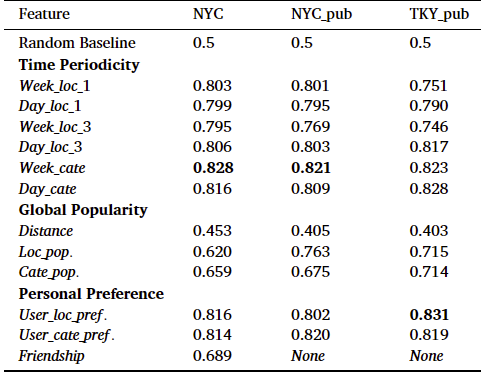
**Accuracy@N：**

分子的意思是表示测试集中命中的次数。

只有当我们在操作列表中对某个位置进行排名时，我们才能成功预测用户在给定时间的未来入住位置。该指标可以解释为成功预测占总预测的比例。

我们将每个用户的签到记录按照9∶1的大小比例按时间顺序进行分割。前90 %的记录用于特征值计算，后10 %的记录用于测试。从这个意义上来说，一定有这样的情况，在测试部分出现的一些未来位置在训练部分不存在。这些位置都不能排序，因为它们不会出现在用户的候选列表中，为了避免这种情况，只有那些在训练和测试数据中出现的位置才使用APR度量来计算。然而，因为我们总是可以判断一个位置是否出现在前N名列表中，所以Accuracy@N可以应用于测试数据中的所有情况。

在三个数据集上进行的第一个实验探索了单个特征的预测能力



**6.监督评分模型和参数推断**

**6.1评分模型**

给定用户的历史签到记录和特定时间，我们可以计算一组特征值。如果我们将所有单独的值视为候选位置的分数，我们可以对所有单独的值求和，并获得每个候选位置的总分数。然而，基于第5节中的分析，不同的特征在位置预测任务中执行不同的任务，因此为不同的特征分配不同的权重是合理的。给定时间候选位置的最终加权评分模型定义如下:



是一个特征向量包含所有的特征的值

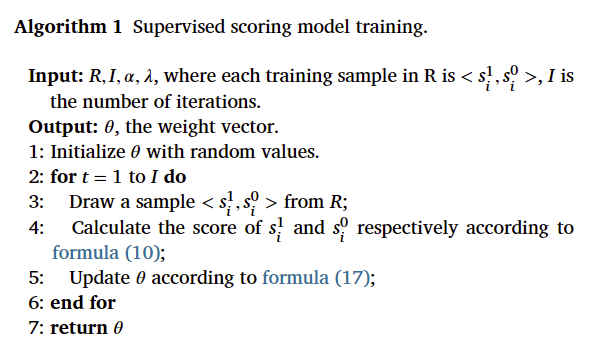
是一个权重向量，对于每一个特征都有一个对应的权重

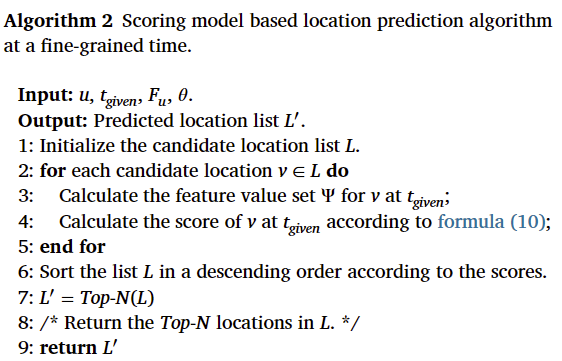
计算出来之后对于候选位置按照从大到小的顺序排列

**6.2参数推断**

每个用户的签到记录按时间顺序被分成9:1两部分，对于一个用户u的前90%的签到记录，我们可以计算他的正例的特征值，也可以计算它的负例（用户在这时没有访问的地点）。我们建立训练模型集合R，每个训练样本是个二元组。这是对于同一个用户的。我们对于参数cita可以使用最大似然估计，我们要保证二元组中前面的数比后面的要大。

利用算法1进行参数推断





**7.监督分类模型：**

针对问题定义2，我们将位置预测问题简化为二进制分类任务，以便对用户是否会在给定时间在给定位置登记进行分类。我们在一个经典的二进制分类框架中处理这个任务。首先，我们解释了训练数据集和测试数据集的构造。所有用户签到记录按照时间顺序按照8 : 1 : 1的大小比例分为三部分D1、D2和D3。对于D1中用户ui的每一个签到记录，如果他在D2中有相应的签到位置(即D2中的相同位置)，我们将通过计算m个特征值来检索正样本，并将该样本分配为正标签+ 1；然后，我们检索D2中u i没有访问过的所有其他登记记录作为负样本，并将这些样本分配给负标签-1。请注意，负样本的选定位置必须在用户ui的候选位置列表中。在这个过程之后，我们可以使用D1中所有用户生成的所有正负样本来构建训练数据集。为了构建测试数据集，我们采取了类似的措施。对于D1+D2中用户u 1的每个签到记录，如果他在D3中有相应的签到位置，我们将正样本标签为+ 1；然后，我们检索D3中ui未访问过的所有其他签到记录，作为标签-1的负样本。通过这种方式，我们构建了测试数据集。

我们采用这种方法来划分训练和测试数据集主要有两个原因:第一，我们希望用户的历史签到记录可以用于预测未来的签到位置，该位置与实际应用中将获得的结果很好地对应；其次，可以保证测试样本独立于训练样本，这对分类模型的评价具有重要意义。在构建训练和测试数据集之后，在训练数据集上训练超可视化二进制分类器，然后在测试数据集上进行评估。值得注意的是，在训练数据集中，正样本和负样本的数量是不平衡的。事实上，负样本的数量远远大于正样本的数量。为了提高分类器对少数样本的灵敏度，我们采用了Chawla中提出的对少数样本进行过采样的方法。

**8.实验以及分析**

由于不同的特征值对应于不同的维度和范围，所以在评分模型和分类模型中都需要归一化，以确保特征值的一致性

**8.1评价指标**

在第5.3节中提到，我们使用APR和Accuracy@N作为监督评分模型的评估指标。

我们使用精确度，召回率以及F1值作为分类模型的评估指标。我们主要使用F1指标来评估分类模型的表现，因为它考虑的更全面。

**8.2对比模型**

1.基本模型：

MFC：定义为全球流行度+用户偏好

MDF：只考虑日模式的特征

MWF：只考虑周模式的特征

MFM：采用一组预测特征，包括用户移动性特征、全局移动性特征和时间特征，以形成监督学习框架

2.一些先进模型

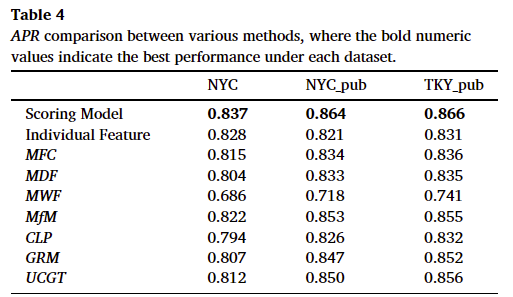
CLP: CLP是一种基于分类器的位置预测方法。机器学习分类器涉及的特征来自三个领域，包括社区、时间和位置流行度。

Grm: 广义回归模型是一种基于地理最近度的高斯混合模型，它利用矩阵分解来预测用户的top-k未来签到位置

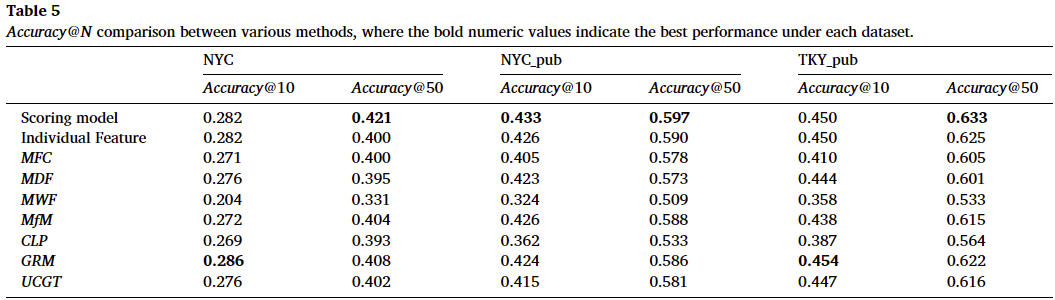
UCGT：是用于用户签到行为预测的集合贝叶斯生成模型。

**8.3结果**

**8.3.1 评分模型**



可以看出来评分模型的APR指标是最好的。



8.3.2分类模型的表现：

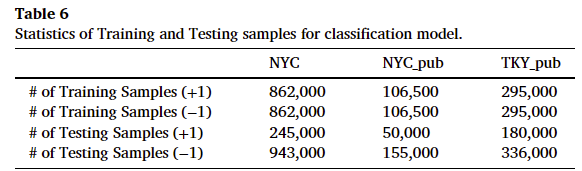
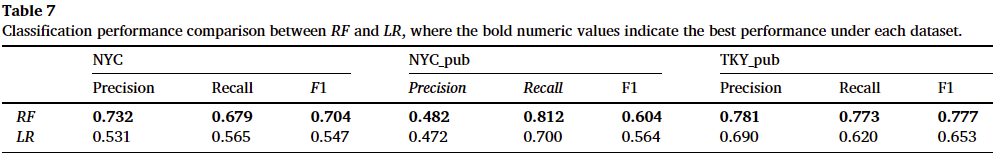
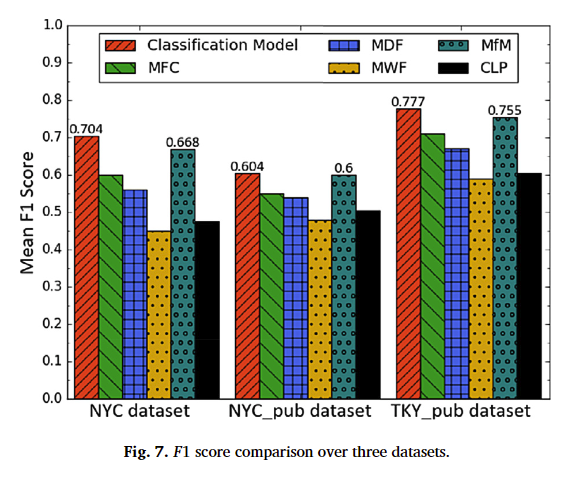


表6总结了分类模型的训练和测试样本的统计

结果如图7所示：





8.3.3将评分模型和分类模型结合

