**基于位置主题的人群分类模型**

**摘要：**先进的手机设备和跟踪技术产生了大量的用户位置数据，通过位置数据能够了解到人们的日常活动的区域特征和位置规律。本文通过网络运营商WIFI信号获取用户日常活动地理位置，从其中研究有相同兴趣爱好、相似生活习惯或工作性质的用户，从而实现人群分类。本文提出了一种基于位置主题的人群分类模型，将用户访问过的地理位置的经纬度作为位置词汇，通过学习建模和概率推导发现位置词汇中隐藏的位置主题及其分布情况，挖掘用户频繁访问的地理位置功能特征，以此作为聚类特征，实现人群分类。实验结果表明，基于位置主题的人群分类模型较现有人群分类方法更能说明用户具有相似的兴趣爱好、生活习惯，并且能过滤离群点和噪音，考虑更加全面，实验效果更优。

**关键词：**位置词汇；位置主题；聚类；人群分类；

## 引言

近年来，手机设备数量的日益增加和跟踪技术的快速发展产生了大量的用户位置数据，这些数据通过各种各样的应用软件记录着人们的日常活动区域，比如说流行的交友软件社交平台陌陌，购物平台淘宝，导航的百度地图等[1]。这些软件记录了用户连接网络运营商WIFI网络时所在的地理位置，这些位置信息反映了人们在真实生活中的活动，而这些活动在一定程度上体现了个人的意图、喜好和行为模式。举个例子来说，如果某个用户经常出现在运动馆，那么可以推断该用户非常喜欢体育运动；更细粒度的分析甚至可以根据用户经常光顾的餐馆类别分析出用户的口味[2]。基于位置数据可以挖掘用户个人行为、生活模式，也可以集合多用户位置数据挖掘人与人的相关性。本文研究后者，通过分析用户历史访问记录，从经纬度坐标中挖掘出地理位置隐藏的语义信息，找出具有相似的兴趣爱好、相似的生活模式或相似的工作性质的用户群体，用于人群分类。人群分类具有很好的应用前景。在很多的系统应用中，通过人群分类，可能发现喜欢同一类书，或者喜欢同一类音乐这类潜在的朋友，并且针对用户群体兴趣点进行推荐。在刑事侦破中，研究犯罪嫌疑人活动范围和经常出现的地理位置，寻找相似活动轨迹的用户，方便确定犯罪同伙目标。由此可见人群分类的重要研究意义。

本文中，基于用户移动设备返回的用户位置信息，对用户常去的区域进行功能划分，并推导出用户访问该功能区域的可能性，以此作为用户位置属性进行人群分类。我们认为：如果用户经常去的区域是属于同一个或同几个功能区域，则认为用户具有相同的兴趣爱好、生活习惯或工作性质，是属于同一类用户。为了更好的说明本文对人群分类的研究，现给出位置词汇和位置主题的定义。

**定义1（位置词汇）**：用户连接网络运营商WIFI信号，并打开上述提到的基于位置服务的APP[1]，由手机发送用户当前地理位置坐标到监测系统，产生的地理位置坐标，即经度和纬度，称为位置词汇。

**定义2（位置主题）**：用户频繁访问的位置词汇是具有一定的功能特性的，比如小区位于生活区域，学校位于教学区域。我们将位置词汇所具有的功能特征称之为位置主题。

在此基础上，本文提出了一种基于位置主题的人群分类模型，不再局限于位置词汇的相同性，更着重发现位置主题的相同性，比如两个用户都喜欢去咖啡厅，虽然两个用户去的是不同的咖啡厅，但是我们认为两个用户具有相同的位置主题，具有相同的生活习惯，属于同类一人群。这种方法着重于发现用户的相同兴趣爱好、生活习惯，由此可见本文的人群分类方法更具有实际相似性，可以有效的提取出相似用户。

## 相关工作

人群分类常常应用在社交网络中，用来在社交网络中寻找有相同兴趣爱好的潜在朋友。在文献[3]中作者采用了最普通的Nearest Neighborhood聚类方法，将Pearson Correlation作为相似性度量标准，计算出特定用户的top-N最相似的用户，从而实现社交网络中的人群分类。文献[5]提出了用户关系的各种信息聚合的方法，但是仅限于用户早就熟识的，并不能计算之前并不认识的用户之间的相似性，在社交网络中不能用于推荐好友。人群分类不仅在虚拟网络中具有良好研究价值，在现实生活中更是如此。文献[6]将人群分类应用在刑事案件侦破中。作者利用主元分析法PCA提取出人群关键行为属性，在获得人群关键行为属性的基础上，利用最大似然估计MLE分类方法对人群进行分类管理。文献[7]针对亚健康状态人群具有的一系列特征，选择logisitic回归建模的方法，确定回归方程，从而得到亚健康状态的判别方程式用于亚健康人群分类。但是上述提到的方法，都是基于用户自身特征属性作为人群分类标准，并没有引入用户位置属性。而位置数据作为虚拟网络与现实世界的唯一链接，并且位置属性是用户行为的重要表述，因此引入位置数据作为人群分类特征是非常有必要的。

文献[8,9]提出根据用户的行为特征可以推断出用户的社会关系：用户喜欢跟朋友或者是相似的用户出现在相同或相似的地方。文献[10]收集三类学生生活的位置数据，采用主成分分析法提取学生特征，使用欧式距离来量化个体间的相似性。文献[11]提出了一个基于层次图相似性度量框架HGSM来度量用户间的相似性。所谓的基于层次图相似性度量，是指首先对用户的位置数据进行层次聚类，划分出不同的地理区域，把用户的历史移动轨迹转化为点与边的序列，并对用户之间的序列进行相似性比较，但该文献采用的是真实地理位置，得出的结果表示两个用户是物理相近。这样的计算方式较简单，但却没有充分挖掘出地理位置的语义信息，对拥有相同位置主题的位置词汇的用户进行人群分类效果并不理想。

而文献[12,13]都是以位置语义信息作为用户相似性判断基准。文献[13]是第一篇将语义替代经纬度来进行用户相似度计算的论文。作者选取其中top-K个用户最常出现的节点，计算用户出现在该top-K中节点的概率，并以此作为两个用户相似性度量的标准。而文献[12]则首先将用户的GPS轨迹数据转化为语义位置信息序列，比如说：商城->餐馆->电影院，然后再采用最大序列匹配算法MTM来计算不同用户间的相似性。

上述将地理位置语义信息应用到人群分类的文献中，都采用停留点判断和位置聚类将地理位置信息抽象为位置语义信息，再针对位置语义信息来进行相似度判断。这类方法的优点是不再将实际的地理位置作为人群分类单位，而是在实际地理位置的基础上抽象出位置语义信息，数据量大大减少的同时也提升了人群分类的层次。但也存在着一些不足，文献[11,12]并没有考虑用户出现在该地理位置的频繁性，也就没有去掉地理位置的噪音和离群点。文献[13]在考虑了用户出现在该地理位置的频繁性，但是需要手动设置参数并其中出现概率高的节点。针对现有人群分类方法的不足，本文提出了一种新的基于位置主题的人群分类模型用于人群分类。该人群分类模型是一种生成模型，通过学习建模和概率推导发现地理位置中隐藏的位置主题及分布情况，并以此作为人群分类特征。这种人群分类方法，不仅考虑了用户文档中位置词汇的实际物理意义，也考虑了位置词汇出现的频率，忽略了噪音和离群点的影响，同时考虑了位置词汇本身所代表的位置主题。就我们的知识，目前学术界还没有相关文献介绍该研究内容。

## 人群分类模型

较现有的人群分类方法，本文提出了新的概念，即位置主题。位置主题作用表现在体现了位置词汇潜在的地区功能特性。文本提出的方法能发现位置词汇的潜在位置主题并计算用户访问各位置主题的概率，发现用户兴趣爱好和生活习惯，并以此作为用户位置属性来进行人群分类。由于本文提出的人群分类模型中的位置主题发现方法是一个概率生成方法，计算出位置主题对应位置词汇出现的概率，忽略了出现概率较低的位置词汇，从而过滤了噪音和离群点，更重要的是，该模型抽取了用户位置文档中位置词汇的位置主题及其分布概率，挖掘出用户的兴趣爱好、生活习惯，将人群分类的用户属性从实际的地理位置转化为用户隐藏在地理位置下的兴趣爱好，并将用户对位置主题的访问概率作为主题聚类方法的用户属性，从而使人群分类结果更具说明性。本文提出的人群分类模型共分为两部分：位置主题发现方法，位置主题聚类方法，整体模型图如图1所示。

### 2.1位置主题发现方法

我们将马尔科夫链的状态转移过程用来拟合位置主题的发现过程，这是由于马尔科夫链的每次状态转移都会产生一个可观测样本并且每个状态都是可达的，可以消除参数不同初始值的影响。我们认为，每个用户位置文档中的位置词汇的位置主题都来自于马尔科夫链中某个状态，其中，位置词汇是每个状态的可观测变量，而位置主题是每个状态的隐藏变量。我们认为，实际地理位置距离不超过200米的位置词汇应该属于同一位置主题，降低了位置主题相互独立的假设。下面是位置主题发现方法的符号约定。

给定用户位置文档，它的位置词汇表示为，表示用户历史访问的地理位置经纬度，其中，是用户位置文档总数。同时使用变量来表示位置词汇的位置主题，比如说商业区、生活区等。假设在给定的用户位置文档集中有个位置主题，则对于每个位置主题。用户位置文档与位置主题的多项分布表示为，并且的狄利克雷先验分布表示为。位置主题与位置词汇的多项分布表示为，并且的狄利克雷先验分布表示为。在每次的状态转移过程中，位置主题的抽取不仅取决于，同时依赖于。如果，会从中为用户文档第个位置词汇重新获取一个新的位置主题，反之，若，则认为用户文档第个位置词汇与前一个位置词汇拥有相同的位置主题。其中，位置主题是隐藏变量，位置词汇是可观测变量。

位置主题发现方法描述如下：

1. 抽取T \* V 多项分布，其中
2. 对于用户文档，其生成过程描述如下：
3. 抽取M \* T多项分布
4. 初始化
5. 对于
6. 如果，，否则，
7. 对于
8. 如果，重新选择位置主题，如果，位置主题

其中，表示与前一位置词汇的实际地理位置距离，表示用户位置文档数，表示位置词汇总数，表示用户位置文档中的位置词汇数目。采用该位置主题发现方法得到的相近距离的位置词汇拥有相同的位置主题，而不再仅仅依赖于，避免了相近位置词汇分配了不同的位置主题，以免对后续的人群分类造成影响。

位置主题发现方法最终要得到是的位置主题及其概率分布结果，即为参数，但是计算的准确值是非常棘手的，通常采用近似推理算法来估计参数。近年的研究中，EM[14]、variational EM[15]、Expectation propagation (EP)[16]以及Monte-Carlo sampling[17,18]都被广泛使用。本文中，考虑位置词汇对于位置主题的后验概率，利用Gibbs抽样间接求得的值，并作为第二部分位置主题聚类方法的输入。首先根据已有参数，表示出联合概率分布如式1所示。



其中，又由和两项构成。当时，取前一位置词汇的位置主题，当时可以进一步表示如式2所示。



    其中， 表示的位置词汇所属的文档，是位置主题k在用户位置文档m中出现的次数，表示当前位置词汇与前一个位置词汇实际地理位置距离超过200米，反之，位置距离小于200米。当时，对积分可得式3。



其中，是狄利克雷分布的统一化系数，表示如式4。



综合以上式子，得到的联合分布如式5所示。



最后根据马尔科夫链状态转换和狄利克雷分布任一元素期望求解性质，可以得到位置主题分布的参数更新公式如式6所示，并将最后求得的作为位置主题聚类方法的输入。



### 2.2位置主题聚类方法

位置主题聚类算法是将每个用户对各个位置主题的访问概率作为权向量，定义为：

,

上式中的表示用户位置文档访问位置话题的概率，其中，且。由于位置主题是位置词汇的语义信息说明，代表着该位置词汇所具备的功能区域特征，如商业区、生活区等。我们认为，用户访问各位置主题的概率可以直接说明用户访问位置主题的倾向性，进一步挖掘出该用户的兴趣爱好和生活模式。比如说，用户如果经常出现在商业区和生活区，那么则是非常典型的上班族生活模式；如果用户经常出现在教学区和生活区，那么该用户可能是学生或教职工。由此可见权向量具备充分说明用户工作性质和兴趣爱好的能力，因此将权向量作为用户的位置属性，作为人群分类的重要标准。

上节中位置主题发现方法得到的位置主题及其分布结果是用户位置文档与位置主题的多项分布，可以表示为。将中各个用户所属的权向量作为位置主题聚类方法的输入，作为用户兴趣爱好和生活模式相似性判断的标准，得到的聚类结果即为人群分类结果。

## 实验

### 3.1 实验数据

在城市大部分区域和农村小部分区域中，会设置移动WIFI信号点。当用户在一个WIFI有效的范围使用基于位置服务的APP时，则会获取用户当前经纬度并通过WIFI信号发送到监测系统中，从而获取到用户的当前地理位置。在实验数据集中，每个被记录下的用户都被赋予了独特的ID，每个用户连接WIFI时都会发送一条位置记录并存储在数据库中。在本研究中，我们搜集了某省某市用户移动手机监测真实数据，共28915848条位置数据，共902906名用户，数据的时间跨度为2015年8月13日到2015年10月10日。详细数据如表2所示，数据格式如表3所示。

表2 数据集描述

Tab 2. The description of data sets

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据集 | 样本个数 | 用户数 | WIFI IP数 | 经纬度数 | 时间跨度 |
| location | 28915848 | 902906 | 73882 | 914030 | 2015-08-13 09:00:02  2015-10-10 05:03:08 |

表3 数据格式说明

Tab 3. The description of data format

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LocationID | Longitude | Latitude | ObjectID | AppType |
| 唯一标识一条记录 | 经度 | 纬度 | 用户ID | App名称 |

进行实验之前，首先对数据进行预处理。

1. 由于位置数据中可能存在着用户位置缓存数据，这些缓存数据是毫无意义的，甚至会有误导作用。例如，一个用户在北京某个地方出差，连接了当地的WIFI，手机上可能就存储着这一位置数据，当回到本地后，再次连接WIFI，就会将在北京缓存的位置数据当做本地位置发送给服务器，但这并不是用户本地位置数据。这类型缓存数据对于实验的研究毫无意义，应该予以剔除。通过百度地图展示获取的经纬度地理位置，会发现某省某市的经度范围为105-106，纬度范围为30-31，其他地区的位置词汇不在本研究范围内。
2. 位置数据集中包含902906个用户移动范围，但却不是每个用户都有很明确的分类。有些用户每个位置词汇出现1-2次，且包含很多位置词汇，对于这种位置主题不明确的用户应予以剔除。

### 3.2评价指标

为了评判人群分类结果的有效性，需要制定正确的评价指标。本文采用聚类算法将用户访问各位置主题的概率作为用户属性进行人群分类。现有的聚类算法指标分为内部评价指标和外部评价指标，内部评价指标用于验证人群分类结果的聚合度，簇内元素相似度越高、簇间元素相似度越低，聚合度越高。现有的内部评价指标包括Davies-Bouldin index、Dunn index和Silhouette coefficient。外部评价指标则采用标签数据或已有的基准来评价。现有的外部评价指标包括Rand measure、F-measure、Jaccard index、Fowlkes-Mallows index、Mutual Information和Confusion matrix。本文采用以下两种评价指标进行人群分类结果判定。

1. 内部评价指标：Dunn index。Dunn index计算公式如下式所示：



表示簇和的距离，表示簇中任意两个元素的距离。因为Dunn index度量的是聚类结果的聚合度，即簇内元素相似度越高、簇间元素相似度越低，聚合度越高。所以值越大，说明聚合度越高，聚类算法效果最好。

1. 外部评价指标：App标签验证。在我们的数据集中，每条数据包含用户通过哪种手机App发送地理位置到监测系统。我们认为，用户当前所处的地理位置的位置主题与使用的手机App存在一定的关系，即如果用户当前处于购物区，使用的手机App很可能是手机淘宝或者京东商城，如果用户当前处于餐饮娱乐区，使用的手机App很可能是百度糯米或者大众点评。基于此认识，将数据字段中的AppType作为人群分类的类标签，采用F-measure验证分类的准确性，计算公式如下：



其中表示准确率，表示召回率，表示和的加权调和平均，是调和参数，当 时，，说明在并不起作用，随着的增大，的影响力也逐渐增大。本 文的人群分类中，设置。

通过对数据集的分析，将人群分5类，分别是商业型、居家教育型、餐饮娱乐型、旅游型、医疗型。其中商业型主要出现在商业区、街道贸易等地区；居家型主要出现在小区和幼儿园、小学附近；旅游型用户主要出现在公园、度假山庄和某省某市著名旅游景点；餐饮娱乐型主要出现在餐馆、小吃甜品点、KTV周围；医疗型主要出现在药店、医院和门诊。

数据集中共包含21种App名称，将其分类为以上5种用户类型标签，如表4所示。

表4 App标签分类

Tab 4. App class label

|  |  |
| --- | --- |
| 类型 | App名称 |
| 商业型 | 陌陌（淘宝），手机淘宝，58同城，京东商城，国美在线，我查查，旺信 |
| 餐饮娱乐型 | 大众点评，百度糯米，美团 |
| 旅游型 | 墨迹天气，百度地图，淘宝酒店，百度旅游，阿里旅行 |
| 居家教育型 | 搜狐新闻，今日头条，新浪微博，暴风影音，腾讯，芒果TV |
| 医疗型 | 今日头条，搜狐新闻 |

### 3.3话题数目选择

采用位置主题发现方法提取位置主题分布时，主题数对位置主题发现方法拟合数据集的性能影响很大，因此需要预先设定主题数。确定主题数的方法有很多种，姚全珠[19]采用贝叶斯统计中的标准方法确定主题数，曹娟[20]通过计算主题主题之间的平均相似度来确定最优模型，Amit Gruber[21]采用困惑度来获得最优主题数。困惑度是衡量一个模型好和坏的评价指标，困惑度越小，代表模型的泛化能力越强，该方法是语言模型中标准的评价标准，在话题模型中被广泛使用。在本次实验中，我们采用困惑度来确定最佳的主题数目。困惑度计算公式如下：



其中表示测试集，，表示用户位置文档d的长度，表示模型产生的概率。

在本次实验中，我们设置，并且。值相对来说较小，这样会让数据集中的词更快的分配到不同的领域的话题中。我们将话题数设置为5,10,15,20,25,30并分别计算困惑度，如图2所示。从图2可以看出，当位置话题数取值为5时，困惑度最高，在位置话题取值为20时到达最低点，之后又有上升趋势，即当话题数目取值为20时效果最好。

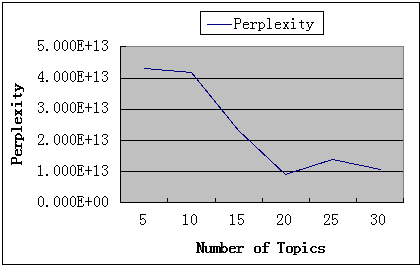


图2不同话题数目困惑度

Fig.2 the perplexity of different number of topics

### 3.4实验结果与分析

位置主题发现方法得到的其中6个位置主题如表5所示，并且取每个位置话题出现概率最大的10个位置词汇用百度地图展示如图3所示，由于位置主题的位置词汇可能分散在不同的地区，图3只展示了各位置主题的部分位置词汇。

表5位置主题

Table5 Position topic semantic representation

|  |  |  |
| --- | --- | --- |
| Topic1 | Topic2 | Topic3 |
| 小区 | 旅游 | 学校 |
| Topic4 | Topic5 | Topic6 |
| 餐饮休闲 | 医院 | 商业 |

Topic1 Topic2 Topic3 Topic4 Topic5 Topic6

图3位置主题

Fig.3 Location word of 6 Location topics

将位置主题发现方法得到的位置主题及用户访问各位置主题的概率作为用户特征，采用位置主题聚类算法实现人群分类。本实验采用现有3种类型的聚类算法进行对比实验，包括K-means，DBSCAN，层次聚类，分别标号为LT-Kmeans，LT-DBSCAN，LT-Agens。同时实现Min-Joong Lee等人[13]提出的另一种判断用户相似性用于人群分类的算法，在文中既引入了地理位置特征，也考虑了用户访问位置的次数，算法记为MPR。采用Dunn index和F-measure两种评价指标评价人群分类结果，实验结果如表6所示。

表5人群分类结果

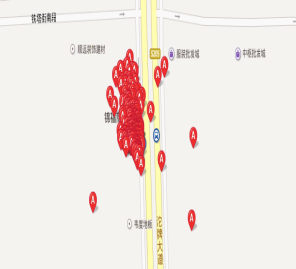
Table5 Crowd classification results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 度量值 | LT-Kmeans | LT-DBSCAN | LT-Agens | MPR |
| Dunn index | 0.585 | 0.302 | 0.563 |  |
| F-measure |  |  |  |  |

接下来将每种分类中的相似用户的位置词汇用百度地图展示观察相似性，如图3-8所示，每幅图从上到下、从左到右依次为LBLTM-Kmeans，LBLTM-Kmedoids，LBLTM-Agens，LBLTM-DBSCAN的实验结果。

Kmeans： #10729 Kmeans：#10107 MPR:#126569 MPR:#18327

Agens：#41053 Agens：#4643 DBSCAN：#132155 DBSCAN：#128495

图3 商业型用户

Fig.3 Commercial users

LBLTM-Kmeans得到的人群分类结果中最相似的两个商业型用户出现位置都包括商业区和住宅区，是两点一线的生活模式；而LBLTM-Kmedoids得到的相似用户都仅出现在商业区；LBLTM-Agens得到的相似商业型用户其一出现在商业区，其二出现在商业区和医院；LBLTM-DBSCAN得到的相似商业型用户其一出现在商业区和小区，其二出现在商业区。由此可见，LBLTM-Kmeans和LBLTM-Kmedoids得到的相似商业型用户效果较好。

综上所述，不管是地图展示结果还是App验证结果，LBLTM-Kmedoids聚类算法较其他三种算法对于人群分类效果更好。

从图中可以看出，MPR找出的同类相似用户虽然都有出现在商业区，但同时有一个用户也访问了餐饮区和小区以及一系列娱乐场所，只能证明用户在一定程度上存在相似性，但相似性不如LBLTM-Kmeans得到的人群分类用户相似性高。同时，MPR需要人为定义用户访问位置词汇频率最高数K的取值，需要定义一系列K的取值通过实验效果来决定最后的取值；本文采用基于位置主题的人群分类方法，采用无监督学习方法，不需要人为标注位置词汇的位置话题，且忽略了出现概率较低的位置词汇。两者之间最大的不同在于，本文提出的方法不仅仅引入位置词汇最为人群分类基准，同时将位置词汇升华为位置主题，以用户经常访问地理位置的功能性来作为人群分类的重要标准，更能说明用户兴趣爱好、生活习惯和工作性质的相似性。

## 结论

本研究在现有的人群分类研究基础上，提出了一种基于位置主题的人群分类模型。相较于传统的人群分类算法来说，本研究不仅考虑了用户访问位置词汇的位置主题，还考虑了用户文档中位置词汇出现的频率，过滤了访问次数较少和访问概率低的位置词汇，而访问位置词汇的频率则从侧面反映了用户的兴趣爱好及生活模式，发现的相似用户更具代表性。在下一步的研究中，将考虑把时间属性加入到人群分类模型中，研究用户在时间维度的位置主题变化，进一步挖掘出用户行为模式及用户移动轨迹相似性。

**参考文献**

[1] ZHENG, Y，ZHOU, X. Computing with Spatial Trajectories. Springer. 2011.

[2]郑宇，谢幸. 基于用户轨迹挖掘的智能位置服务[J].中国计算机学会通讯，2010.

[3]B Sarwar,G Karypis,J Riedl.Application of Dimensionality Reduction Recommender System--A Case Study[J].2000.

[4]Ellen Spertus,Mehran Sahami. Evaluating similarity measures: a large-scale study in the Orkut social network[J].ACM.2005,August 21-24.

[5]Mt.Carmel Haifa,San Jose.Do you know?: recommending people to invite into your social network. In: International Conference on Intelligent User Interfaces[J].ACM.2009.February 2-11.

[6]张成，刘亚东，谢彦红等.基于PCA与MLE方法的人群分类新方法研究[J].沈阳化工大学学报.2015-06-15.

[7]刘保延，何丽云等.亚健康人群分类及其临床特征分析与评价[J].世界科学技术.2006-04-30.

[8]Jacob Goldenberg,Moshe Levy.Distance is not dead: Social interaction and geographical distance in the internet era[J].arXiv preprint arXiv:0906.3202,2009.

[9]Diana Mok,Juan-Antonio Carrasco,Barry Wellman.Does distance matter in the age of the internet[J].Urban Studies,2010.

[10]宋衡.基于位置数据的人类行为识别和相似性研究[D].上海交通大学.2014-02-16.

[11] Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., Ma, W.Y.: Mining user similarity based on location history. In: 16th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, p. 34 (2008)

[12]Xiangve Xiao,Yu Zheng,Qiong Luo,Xing Xie.Finding Similar Users Using Category-Based Location History[R].18th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2010, November 3-5, 2010, San Jose, CA, USA, Proceedings.

[13]Min-Joong Lee,Chin-Wan Chung.A User Similarity Calculation Based on the Location for Social Network Services[J].DASFAA 2011, Part I, LNCS 6587, pp. 38–52, 2011. c Springer-Verlag Berlin Heidelberg 2011.

[14]Thomas Hofmann. Probabilistic latent semanticanalysis. In Proceedings of the 15th Annual Conference onUncertainty in Artificial Intelligence(UAI-99), pages 289–29, San Francisco, CA, 1999.Morgan Kaufmann.

[15]David M. Blei, Andrew Y. Ng, and Michael I.Jordan. Latent Dirichlet allocation. Journal of  
Machine Learning Research, 3:993–1022, 2003.

[16]Thomas Minka and John Lafferty. Expectationpropagation for the generative aspect model. InProceedings of the 18th Conference on Uncertainty in Artificial Intelligence, pages 352–359,San Francisco, CA, 2002. Morgan Kaufmann Publishers.

[17]Michal Rosen-Zvi, Tom Griffith, Mark Steyvers,and Padhraic Smyth. The author-topic model for  
authors and documents. In Max Chickering andJoseph Halpern, editors, Proceedings 20th Conference on Uncertainty in Artificial Intelligence,pages 487–494, San Francisco, CA, 2004. MorgamKaufmann.

[18]Thomas L. Griffiths, Mark Steyvers, David M.Blei, and Joshua B. Tenenbaum. Integrating topics and syntax. In Lawrence K. Saul, Yair Weiss,and L´eon Bottou, editors, Advances in Neural Information Processing Systems 17, pages 537–544.MIT Press, Cambridge, MA, 2005.

[19] 姚全珠，宋志理，彭程. 基于LDA模型的文本分析研究[J]. 计算机工程与应用，2011,47(13)：150-153.1002-8331(2011)13-0150-04.

[20] 曹娟，张勇东，李锦涛等. 一种基于密度的自适应最优LDA模型选择方法[J]. 计算机学报， 2008，31(30)：1780-1787.

[21]Amit Gruber, Michal Rosen-Zvi,Yair Weiss. Hidden Topic Markov Models[J].2007.