移动社交网络与用户位置

谢　幸[1]　连德富[2]
[1]微软亚洲研究院
[2]中国科学技术大学

关键词：空间数据挖掘　位置服务

背景

在互联网上流传过这样一个笑话：“每个保安都是哲学家，他们每天都在提出哲学中的三个终极问题：你是谁？你从哪里来？你到哪里去？”

实际上，这也是三个很好的研究问题。我们可以换种问法：你是什么样的人？——由此推断你偏好哪种类型的地方；你去过哪些地方？你又将会到哪些地方去？这三个问题的解答可以帮助我们提供更加个性化的服务。举个例子，如果你现在正在某个商场购物，而且购物后一般会选择附近的快餐馆去用餐，那我们知道后就可以提前将你可能会感兴趣的餐馆的优惠券发到你的手机上。同时，这些信息对于在线广告、移动推荐和很多其它商业应用都有着巨大的价值。

俗话说，巧妇难为无米之炊。为了解答上述问题，我们首先需要拥有足够的用户数据，这需要获取并连续记录用户曾经去过的地方。近年来，全球卫星定位系统（GPS）、基站定位和各种室内定位技术迅速发展，人们已经拥有了非常成熟便捷的位置获取技术。据高德纳（Gartner）咨询公司预测，到2013年将会有85%的移动设备预装GPS。4年前，我们曾经为本刊撰写了《基于地理信息的用户行为理解》一文[1]，文中比较了常见的定位技术，并介绍了一些相关的研究工作。

尽管现在获取用户位置已不再是难事，但是大规模的搜集用户位置数据却仍然存在着挑战。由于用户的位置数据可能存在于各种类型的数据中，可以是带地理标注的照片、微博和游记，也可以是位置搜索日志和地图服务日志等。目前我们仍缺乏一个很好的机制能够集中管理这些来自不同设备、不同服务和不同用户的位置数据。

然而，移动社交网络的发展给这个问题的解决带来了新的契机。在社交网络中，用户主动和他们的朋友们分享心情、爱好、活动和照片等各种信息。这些信息大多显式或隐式地包含了用户的位置。最近兴起的基于位置的社交网络（location based social networking，LBSN，或称为签到服务）允许移动用户在社交网络中共享各自的位置以及与位置相关的信息。例如，用户在餐厅吃饭时，可以通过手机定位，签到他所在的餐厅名称并分享给他的好友。在分享时，用户还可以附加上自己对该餐厅及菜品的评价甚至照片。图1就是这样一个实际的例子。目前提供签到服务比较知名的有Foursquare、街旁、开开和切客等网站。以Foursquare为例，它从创立到达到100万用户，只花了短短一年时间，比Twitter少用了一半时间。截至2012年4月，Foursquare在全球的用户数目就已经超过了2000万。从Foursquare成立至今，用户签到的总次数已超过了20亿次，这是一个非常庞大的用户位置数据集。一方面，这些签到服务鼓励了移动用户和他们的朋友通过共享位置信息的方式保持联系；另一方面，这些服务也让我们可以用一种相对方便的方式来搜集并集中管理用户的历史位置信息。可以预测，在未来几年内，单纯的位置签到功

图1 位置信息共享实例

能将会逐渐被主流社交网络服务吸收，并与其它功能相融合，比如照片标注、餐馆点评、游记共享和周边朋友搜索等。近期，基于位置的社交网络被列入“十二五”国家科技计划备选项目库（地球观测与导航技术领域），也间接说明这个领域已开始受到学术界的广泛重视。

下面将着重介绍近期我们在基于移动社交网络数据以理解用户位置方面开展的研究。

用户位置命名

我们关心的第一个研究问题是“用户在哪儿？”

这个问题貌似很简单。前面也谈到，定位技术的发展使得获取用户位置不再是件难事。但是GPS坐标只是一些数字，并不能直接反映用户实际所在的位置名称。例如，当看到坐标（31.2398, 121.4997）时，你能猜到它代表的是东方明珠塔吗？以Foursquare为例，虽然它可以通过移动设备定位用户，但在签到时，还是需要用户直接输入位置名称，或是从一个附近的位置名称列表中选择他所在的位置。这主要是由于GPS坐标和位置名称对应关系不唯一所造成的。

有人会问，为什么需要知道用户确切的位置名称呢？因为对很多位置服务来说，位置名称可用于对结果的排序和过滤。举例来说，如果我们确切地知道某用户是在一家川菜馆，就可以向他推荐热门或美味的川菜，而不是上海菜或者其它。另一个应用的例子是用户轨迹数据管理系统。这些系统通常只是在地图上显示轨迹，却并不附加任何语义标注。实际上，人们往往不久就会忘记他们过去曾在这些地方做了些什么，甚至会忘记这些地点的名称。根据这种情况，我们可以开发一个自动的位置标注工具，以方便用户理解他们所积累的轨迹数据。类似的应用还有很多，这里不一一列举。

用图2来示意要研究的问题，其中A、B和C为三个用户。他们的地理坐标非常接近。由于定位的误差，只知道他们都在同一个小区域内，但并不知道他们各自的具体位置。已知的情况是：用户A在附近上班，经常在其所在的公司签到；用户B是一个“空中飞人”，经常穿梭于各城市之间并在下榻的旅馆签到，这是他第一次来到当前地点；用户C经常在娱乐场所签到，来这里很有可能是想找个地方和朋友聚会。因此，基于他们不同的签到历史，有理由认为用户A正在他的公司工作；用户B在该区域内的某家旅馆居住；用户C则应该是在某个酒吧和朋友在一起。这个例子生动地说明了GPS坐标和位置名称不是一个简单的对应关系，我们需要根据用户的不同特性来推测用户真正所处的位置。

过去有很多研究工作试图从个人的GPS轨迹数据中抽取对他们有重要意义的地点数据。例如文献[10]对用户停留超过10分钟的地点进行聚类来找到重要地点，并使用二阶马尔可夫模型对用户在相邻地点之间的转移行为进行建模。文献[8]在原始GPS数据中使用层次马尔可夫模型抽取重要地点，并检测用户的交通模式。这些工作主要着重于在轨迹数据中找到重要的区域，但并不为这些重要区域标上有意义的名称。

近期也有一些研究工作开始为抽取的地点进行标注。这些工作通常是先从轨迹中抽取重要地点，然后根据情境信息和历史数据对它们进行标注。文献[4]使用关系马尔可夫网为地点标注用户在该地点进行的活动。文献[5]使用层叠条件随机场同时抽取

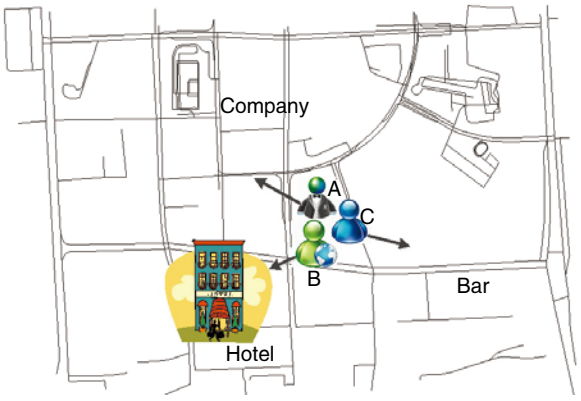


图2 用户位置命名问题

并标注重要地点。但在他们的标注集合中只有四类名称，即公司、家庭、朋友家和停车场。他们的算法也很难增加大量新的地点名称。文献[6]研究了当用户和朋友共享位置信息时，对位置命名方式的偏好；研究了位置的访问频度熵、社交网络、朋友对地点的熟悉程度和分享人对地点的隐私程度等影响因素。但该工作将位置命名偏好只分为街道地址、语义地址和混合地址三类，并不试图得出具体的名称。我们则希望能标注具体的位置名称，即在兴趣点（point of interest，POI）层次上进行标注。我们的工作是在这些工作的基础上往前迈了一大步。

问题分析

解决位置命名问题的一个直观方法就是使用地图服务提供的地址逆查功能。地址逆查功能可以将GPS坐标转换为相应的街道地址、邮政编码或者行政区划名称。但是正如文献[6]指出的，用户很少在社交网络中使用街道地址（例如，北京市海淀区丹棱街5号）来描述他们的位置。而同一个街道地址也经常会对应到多家不同的商店和餐馆，例如他们都集中在同一个大型综合商城里。更重要的是，很多应用都需要有意义的地点名称，尤其是功能性的描述，如前面提到的餐馆、商店和酒吧等。因此，我们需要将地理坐标映射到兴趣点名称。

实现这个映射不是一件简单的事情，主要难点在于坐标的误差，兴趣点的高密度、多层次以及大尺度特性。

两年前，我们曾在本刊组织了一期“基于位置的服务”专题。在该专辑中，来自武汉大学测绘学院的作者指出，GPS定位通常有30米左右的误差，而基站定位的误差可能会达到100米甚至更大[2]。尽管GPS定位有较高的精度，但它非常依赖环境状态，仅仅是天空的遮挡程度也会对定位精度有很大的影响。实际应用中，很多用户签到的地点都没有很好的GPS信号。

基于地图数据，我们对北京所有的兴趣点进行了统计，发现在边长100米的正方形区域内平均有6个兴趣点，最多的时候有286个兴趣点。在这种高密度区域中，将GPS坐标映射到兴趣点名称会十分困难。兴趣点也具有层次性，例如同样在北京市海淀区丹棱街5号，用户可以签到微软大厦、微软中国研发集团、微软亚洲研究院，甚至某个具体的会议室。另外，有些类型的兴趣点的覆盖范围会相当大，比如清华大学，如果简单地计算用户所在位置与清华大学中心位置的距离，以此来推测用户的位置名称，其结果往往是不准确的。兴趣点的这些特征也给位置命名问题带来了难度。

解决这些问题的基本出发点是依靠用户的历史数据。通过观察，我们发现位置命名和位置搜索有很多相似性：二者的输入都包含用户的地理坐标，输出就是一个排序的兴趣点列表。它们的不同点是用户在进行位置搜索时需要输入查询词，以表达他们具体的信息需求。根据查询词和兴趣点之间的相似性以及用户和兴趣点之间的距离，位置搜索对结果进行排序，而在位置命名问题中，用户没有任何人工输入。这里用户本身的历史信息更加重要，假设用户一般会遵循以往的规律进行活动，如图2所示，如果不知道用户的身份，将很难给出准确的结果。实际上，我们的系统试图在没有任何人工干预的情况下猜测用户所在的确切位置，位置搜索系统基于用户给定的关键词找到用户心目中的位置，而不是其所在位置。

算法设计

基于上述讨论，参照位置搜索系统的方法来解

决位置命名问题。首先进行离线的特征抽取，从互联网、评论网站和位置服务等来源中抽取兴趣点的静态特征，通常代表它们的热门程度；然后，结合兴趣点和用户当前情境信息之间的动态特征，使用排序学习（learning to rank）的方法来建立模型。

我们的系统使用了静态和动态两类特征。兴趣点的热门程度特征通常是静态的，与用户当前的状态无关。我们从互联网和位置服务中抽取这些特征。例如，通过搜索引擎查看兴趣点在互联网中出现的频度，频度越高代表越热门。最近生活点评类网站，如大众点评网，非常流行。用户在这些网站上对餐馆、酒吧以及其它类型的兴趣点进行点评；从一些点评网站上抽取三种特征：点评的数目、平均点评分数和照片数目。当用户使用位置服务时，他们的行为会被系统记录下来。这些行为代表了他们对兴趣点的关注，可以作为兴趣点的热门程度因素之一。我们还考虑了两类信息来源：基于位置的社交网络和位置搜索。在基于位置的社交网络，例如在Foursquare中，有更多签到数目的地点比那些签到数少的地点要更加热门。我们同时计算签到的数目和签到的人数，在位置搜索中，访问次数更多的兴趣点显然也更加热门。在实验中，使用了必应地图服务的日志来计算兴趣点的访问次数。

动态特征和用户的当前状态相关，代表用户和兴趣点之间的联系。给定一个用户以及当前的地理坐标和时间，我们为数据库中每个兴趣点抽取下面几种特征：

用户地理坐标和兴趣点之间的联系 使用经纬度来表示兴趣点和用户位置，从而可以计算它们之间的距离。

当前时间和兴趣点之间的联系 实际上，当前时间是推测用户位置很有用的信息。例如，为了节约时间，用户访问快餐店的概率在工作日比周末要大。我们在算法中将时间按三种粒度，即一星期中的第几天，工作日还是周末，以及一天中的第几个时间窗口（以4小时为单位），进行量化。

用户签到历史和兴趣点之间的联系 根据用户的签到历史来计算用户访问一个具体兴趣点的频度。这个特征可看成兴趣点相对于某用户的热门程度。

事件特征 根据时间和地点可以唯一确定当时发生的事件。很多事件会对兴趣点的计算会产生影响。例如，如果当时正在下雨，用户所在的位置就会和晴天不同。其它可能的影响因素包括商店的开张、关闭或减价等不同条件下的计算结果也会出现不同信息。

获取了这些不同种类的特征后，需要考虑如何结合它们得到最终结果。我们使用了在信息检索领域中的排序学习算法（更多细节请参考文献[7]）。对于用户签到历史中的每个记录，用户签到的兴趣点标注为相关，数据库中其它附近的兴趣点则标注为不相关。

实验

我们的实验数据集来自大众点评网(www.dianping.com)。大众点评网不仅提供兴趣点的点评功能，还允许移动用户签到和共享他们的位置。每个签到包括用户的位置、签到时间、兴趣点的主页和用户的评价信息。我们抓取了2011年1月7日到6月11日之间来自北京的用户的签到数据，统计结果如表1所示。

表1 大众点评网数据集描述

兴趣点总数目	15664
用户总数目	545
签到总数目	31811
平均每个兴趣点的签到数目	2.6
平均每个兴趣点的用户数目	1.4
平均每个用户的签到数目	58
平均每个用户签到过的兴趣点数目	32
时间跨度（天）	152

经研究发现，部分用户平均每天提交20次甚至30次签到。通过分析他们的记录，我们发现其中一些人在很短时间内多次签到同一个地点，还有一些人纯粹为了好玩或是为了获取优惠在短期内签到多个不同的地点。我们认为后者属于恶意签到用户。这些用户产生的记录对位置服务以及相关的数据分析都会造成恶劣的影响。因此，我们在实验中设计了一个简单的过滤方法，通过查看平均每天签到次数和相邻签到的时间间隔来过滤这些恶意用户。这个过滤方法在实验中

的效果很好。当然，恶意签到用户的检测本身也是一个很有意义的研究问题，我们鼓励读者去思考并设计一个更加有效的检测算法。

此外，我们还发现用户签到历史、距离、评论次数和网页出现频度都是重要的特征。实验的数据集上输出的前5个签到结果可以达到64.5%的成功率。具体内容可参见文献[3]。

用户位置预测

关注的第二个问题是“用户要去哪儿？”

用户位置预测和位置推荐是两个相关的问题。位置预测着重在不干扰用户正常行为的前提下预测用户未来的位置，而位置推荐强调的是推荐结果的新颖性，希望为用户推荐他可能会感兴趣而以前未去过的地点，从而改变用户的行为。文献[9]设计了一个基于贝叶斯学习的系统，通过用户的偏好和位置情境信息来推荐餐馆。此外，还有很多工作使用了协同过滤来做餐馆或其它类型的地点推荐。这里不准备介绍具体的算法，就把这个问题留给读者。下面主要针对使用签到数据来做用户位置预测并进行一些初步分析，供大家参考。

直观地看，人们会想到用户访问位置之间的马尔可夫性质。一般来说用户当前的位置对他们计划要去的位置有很大影响。例如，在用户上午签到了公园以后，他们一般会去吃午饭，所以很可能会在附近的餐馆签到。

我们在数据集里发现，人们经常在一段时间内签到多于一个地点，连续两次签到之间的间隔有50%以上是在8分钟以内，它们之间的距离小于256米的概率也大于50%。很多签到都是在一个小区域内以及短的时间段内发生的。这些连续签到之间未必有严格的顺序关系，也就不能简单地使用马尔可夫模型来建模。此外，签到的时空模式也很重要。用户一般倾向于在和上一次签到位置不远的地方签到，相邻签到之间的距离大多都不超过4公里，距离越长，签到概率衰减得也越快，这主要是因为用户在短时间内不会移动很长的距离。用户的签到模式也具有周期性，不管是周末还是工作日，用户中午都会倾向在餐馆签到。有时，用户签到兴趣点的类型也具有一定的动态性，周末的下午，用户可能会有更多运动和娱乐方面的签到，而在工作日，这类签到可能会发生在晚上。

总之，在设计位置预测算法时，需要考虑到数据的稀疏性、签到的长期顺序性和短期非顺序性以及签到的时空模式等多种因素。

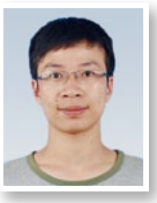
结语

我们探讨了移动社交网络对用户位置理解和预测等问题的影响，并介绍了一个用户位置命名算法。基于用户的地理坐标、当前时间以及签到历史，我们为用户的当前位置提供具体而有意义的名称；我们借鉴位置搜索和位置命名的相似性，提出了一个参照位置搜索的算法框架，结合热门度特征和用户历史特征，来解决位置命名问题。

未来，我们准备通过聚类相似用户或者协同过滤的方法来解决数据的稀疏性；还将考虑更多类型的特征，尤其是与用户相关的特征。由于人们对隐私保护越来越关注，为此我们将考虑在兴趣点的类别层次上进行位置理解和预测，并且在一定程度上也可解决数据稀疏性的问题。■

谢　幸

CCF高级会员、本刊编委。微软亚洲研究院主管研究员。主要研究方向为空间数据挖掘、基于位置的服务、社交网络和普适计算等。xingx@microsoft.com



连德富

CCF学生会员。中国科学技术大学博士研究生。主要研究方向为空间数据挖掘和移动推荐等。

liandefu@mail.ustc.edu.cn

参考文献

[1] 谢幸, 郑宇. 基于地理信息的用户行为理解. 中国计算机学会通讯, 4(10), 2008

[2] 李清泉, 乐阳. 基于位置服务的分析与展望. 中国计算机学会通讯, 6(6), 2010
[3] Lian, D., Xie, X. Learning Location naming from user check-In histories. In Proc. ACM GIS 2011, Chicago, IL, USA, Nov. 2011
[4] Liao, L., Fox, D. and Kautz, H. Location-based activity recognition using relational Markov networks. In Proc. IJCAI 2005, 773 ~ 778
[5] Liao, L., Fox, D. and Kautz, H. Extracting places and activities from gps traces using hierarchical conditional random fields. The International Journal of Robotics Research, 2007, 26(1):119 ~ 134
[6] Lin, J.L., Xiang, G., Hong, J.I. and Sadeh, N. Modeling people's place naming preferences in location sharing. In Proc. UbiComp 2010, 75 ~ 84
[7] Liu, T.Y. Learning to Rank for Information Retrieval. Foundation and Trends on Information Retrieval, 2009
[8] Liao, L., Patterson, D.J., Fox, D. and Kautz, H. Learning and inferring transportation routines. Artificial Intelligence, 2007, 171(5-6): 311 ~ 331
[9] Park, M., Hong, J. and Cho, S. Location-based recommendation system using Bayesian users preference model in mobile devices. In Proc. Ubiquitous Intelligence and Computing 2007, 1130 ~ 1139
[10] Ashbrook, D. and Starner, T. Using GPS to learn significant locations and predict movement across multiple users. Personal and Ubiquitous Computing, 2003, 7(5):275 ~ 286

[2] 吕金虎,王红春,何克清,复杂动力网络及其在软件工程中的应用,计算机研究与发展,2008,45(12):2052 ~ 2059
[3] J. Lü,X. Yu,G. Chen, and D. Cheng, Characterizing the Synchronizability of small-world dynamical networks, IEEE Trans. Circuits Syst. I, 2004, 51(4): 787 ~ 796
[4] J. Lü and G. Chen, A time-varying complex dynamical network model and its controlled synchronization criteria, IEEE Trans. Automat. Contr., 2005, 50(6): 841 ~ 846
[5] J. Zhou, J. A. Lu, and J. Lü, Pinning adaptive synchronization of a general complex dynamical network, Automatica, 2008, 44: 996 ~ 1003
[6] M. Cha, J. A. N. Perez, and H. Haddadi, Flash floods and ripples: The spread of media content through the blogosphere, Proc. 3rd Int. AAAI Conf. Weblogs Social Media, San Jose, California, USA, May 2009
[7] M. Cha, A. Mislove, and K. P. Gummadi, A measurement-driven analysis of information propagation in the flickr social network, Proc. the 18th Int. World Wide Web Conf., Madrid, Spain, 2009: 721-730
[8] D. Centola, The spread of behavior in an online social network experiment, Science, 2010, 329(5996): 1194 ~ 1197
[9] X. Song, Y. Chi, K. Hino, and B. Tseng, Identifying opinion leaders in the blogosphere, Proc. the 6th ACM Conf. Conf. Inf. Knowledge Manag., Lisboa, Portugal, 2007, 971 ~ 974
[10]A. Anagnostopoulos, R. Kumar, and M. Mahdian, Influence and correlation in social networks, Proc. the 14th ACM SIGKDD Int. Conf. Knowledge Discov. Data Min. (KDD), ACM, New York, 2008, 7 ~ 15