基于地理信息的用户行为理解

谢幸，郑宇

用户行为理解

用户行为理解对于提供个性化互联网服务和广告有着极为重要的意义。目前的商用或研究性系统试图通过分析用户的在线行为以理解用户。它们观察用户如何进行网络搜索，在互联网上阅读以及写作，通过研究这些在线活动的规律以学习用户的兴趣喜好，例如，我们可以从用户阅读历史中发现他们对哪一类新闻更加关心，从而设计更好的个性化新闻阅读体验。然而，这些方法忽略了用户的非在线行为，或称为”物理”行为。这里的物理行为指的是人们是如何进行购物，用餐，旅游和其它任何发生在真实世界(或称为”物理世界”)的活动。相对于在线行为，这些活动往往更能准确和完整的代表用户的兴趣。

在本文中，我们关心的是这些活动中的地理位置信息。这里的地理位置信息是指用户活动的场所，通常用经纬度坐标，地名或兴趣点名称等来表示。地理位置是人们日常行为中最重要的因素之一，包含着大量的信息。随着移动定位设备和电子地图等互联网地理信息工具的快速发展，检测用户物理行为中的地理位置信息变得更加方便，因为人们的物理行为往往会通过在线行为反映出来。例如，人们通常会使用搜索引擎来规划自己未来的活动，比如在和朋友聚会前搜索一个评价比较好而且离家不远的海鲜餐馆；研究一条更快的去往目的地的行车路线；以及在网上阅读有关该餐馆的评价文章。在旅游时，不少用户通过使用支持全球卫星定位系统(GPS)的移动设备来记录他们的位置轨迹，并将它们在网上共享给自己的朋友。另外，还有很多人通过上载博客文章或者照片来和朋友分享他们的快乐。在这些应用场景中，地理位置信息可以通过分析搜索查询词，GPS 日志，博客文章甚至图片来获取。

基于不同类型的数据，我们可以挖掘不同的用户兴趣。例如，搜索查询词包含了用户感兴趣的地点信息，而个人的位置轨迹可以部分的反映用户的生活规律，博客文章和照片则可以给出有关特定地点更丰富的信息。通过分析多人的数据，我们还可以进一步了解和地点相关的统计特性以及不同地点和用户之间联系。在本文中，我们将重点介绍针对两类数据，即搜索查询日志和个人位置轨迹的研究工作。

理解用户轨迹

随着无线定位技术的高速发展，人们逐渐拥有了快速获取自己当前位置的手段。无论是全球卫星定位系统，还是基于无线蜂窝网的手机定位技术，都让人们能更高效地认知周边地理环境。这些位置信息不但可用于定位、导航以及提供一些基于位置的服务，也可用于表达用户在地理空间的历史行为。将一个用户孤立的位置点按照时间顺序连成线路，便可表

达该用户过去的历史轨迹。多条历史轨迹的累积便可用来反映用户的生活规律和行为特征。而从大量用户的数据集合中则可分析出一个区域内人们的生活模式和社会规律，如热点地区、经典旅行线路和交通状况等。

在当前众多的无线定位技术中，GPS 以其覆盖范围广、定位精度高、定位时间短和定位依赖性小等优势逐渐在人们的日常生活中变得普及起来（见表 1）。各种车载 GPS、手持 GPS 和 GPS 智能手机的相继问世也为人们提供了更加便捷的位置获取和轨迹记录方式。作为用户经历的载体，这些轨迹数据在各种应用中发挥着重要的作用，并帮助人们来理解个人行为和社会规律。从数据源来看，当前的研究工作可分为基于个人轨迹数据的理解和基于多人轨迹数据的理解两个方向。

表 1：不同的位置检测技术及其比较

	卫星定位	蜂窝无线定位	无线局域网定位	RF 定位
适用环境	室外	室内、外	室内	室内
定位精度	高	较低	中	高
普及率	较高	高	较低	较低
定位时间	快	较长	较长	快
覆盖范围	广	较广	较小	小

理解个人数据

由于位置检测技术的迅猛发展，用户可在不干扰生活的前提下轻松地记录自己的旅行线路、运动经历、以及日常生活和工作轨迹。结合现有的地理信息数据库和电子地图，这些轨迹数据可为个人提供以下服务。

- 帮助用户更有效的回忆过去：个人的轨迹数据可看作是一种自动化的电子日记，从中用户可以清楚地了解自己过去的经历。比如，从这些数据中用户可以准确的知道上星期五自己的上班时间，午餐就餐地点以及在回家路上花费的时间等信息。这种功能对于外出旅行和户外运动更加有效。

- 更便捷的与朋友分享生活经历：互联网的普及催生了网络博客的发展。通过博客，朋友之间可以方便的分享近期的生活经历。最近在互联网上出现了一种以 GPS 轨迹数据为中心的新兴应用。在这些互联网的虚拟社区里，用户可以通过发布自己的轨迹数据来展现自己的旅行经历或运动线路[1, 2, 3]。比如，自行车爱好者可以将自己的骑行线路利用 GPS 设备记录下来，然后通过互联网上载到论坛来与其他爱好者交流和分享。

- 理解自己的生活规律，提供个性化服务：当个人的数据积累到一定程度，该用户的生活规律已经在数据中得到了体现。因此，相当一部分的研究工作从个人的长期数据中分析出对用户具有重要意义的地点，比如家、公司和常去的商场和餐厅。进一步，根据用户过去的经历得出用户在这些地点的转移概率，从而能够对用户今后的活动作出

较为准确的预测。例如，当用户被预测出将要前往某个商场，系统可将该商场的促销信息提前发送到用户的手机上。

理解多人轨迹数据

单个用户的轨迹数据可以体现个人的生活规律，而多个用户轨迹数据的集合则可用来表达一个社区乃至城市里人们的生活模式。

- 热点地区和经典线路检测：基于大量用户的数据，我们可以检测一个城市里的热点地区及经典的旅行线路。这些信息可以帮助人们快速了解一个陌生的城市，并为游客推荐风景名胜和出行线路。例如，从数据中可以发现，前来北京天安门游览的游客通常会参观附近的故宫。此后，有相当一部分的人会继续去后海游玩。这条经典的旅行模式可为其他游客提供先验知识，从而高效地观光更多的景点。

- 交通状况分析和预测 [4, 5]：根据多个用户在城市道路上记录的轨迹，我们可以分析和预测不同路段的交通状况。与实时路况相比，此类信息可提前预告交通状况，为用户出行的路线规划提供参考，避免因交通堵塞而造成的时间浪费。

- 用户行为识别 [6, 7, 8]：用户行为包括在固定目的地的行为，如就餐、购物、运动等，也包含对用户在路程中的行为理解，例如用户当时采用的交通方式是开车、公交、自行车等以及预测用户可能选择的目的地。该类研究方向的特点是从多个用户的历史数据中提取出特征，然后利用这些特征以监督学习的方法训练出一个分类模型。日后，当遇到其他用户的轨迹数据时，该分类模型可自动识别他人的行为。

GeoLife：个人轨迹数据管理系统

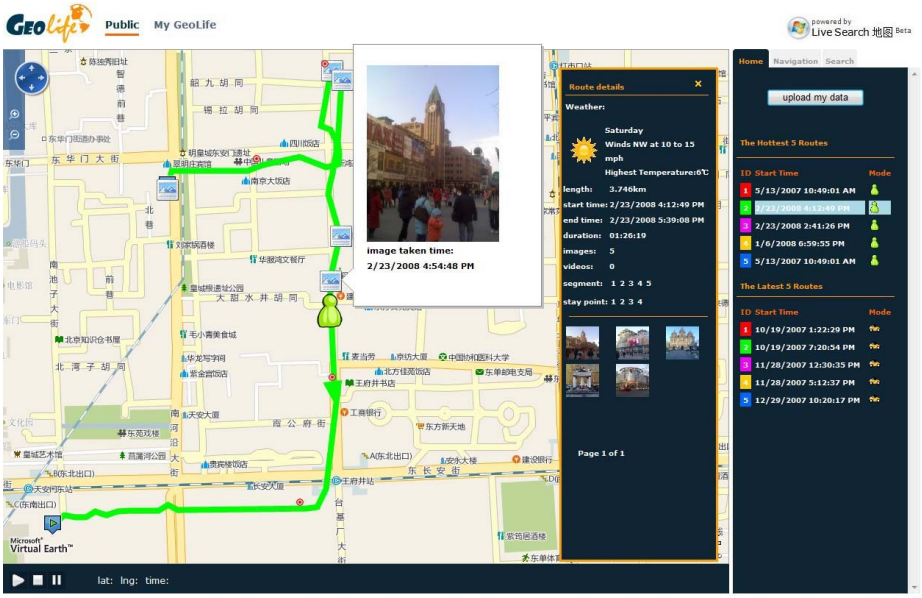


图 1：GeoLife 系统界面

GeoLife（地理人生）[9, 10]是我们研发的一种以 GPS 数据为中心并基于电子地图的应用系统（见图 1）。它不但是可视化、管理和理解个人 GPS 数据的工具，也是多个用户共享 GPS 数据和交流生活经历的平台。基于个人的 GPS 数据以及相关联的多媒体内容，GeoLife 在地图上以动画的形式生动地重现用户过去的经历。这不但有助于用户有效地回忆自己过去的往事，也可成为一种朋友之间交流生活经历的更便捷、更直观的方式。从不断累积的个人数据上，GeoLife 还可帮助用户了解自己的生活规律，以保障健康的生活习惯。当更多的用户将自己的数据发布到公共平台与朋友分享后，人们不但可以从其他人的数据中借鉴经验和分享快乐，也可以发现热点地区和经典的旅游线路等社会群体规律。

轨迹中的交通方式判别

理解用户的交通方式是 GeoLife 系统中的核心模块之一。用户只需要提交记录的 GPS 轨迹数据，系统就能自动推断出用户当时采用的交通方式是驾车、骑自行车、公交车或步行。该技术的难点在于以下两点：

- 用户在一次行程中通常会采用两种以上的交通方式。例如先步行到公交车站，然后搭公交车前往目的地，最后再通过步行在目的地活动。
- 用户的移动速度容易受到交通状况和天气因素的影响。在拥堵的交通中，用户的驾车速度甚至不如自行车的骑行速度。因此，纯粹基于速度的判决方法不能准确地区分不同交通方式。

交通方式的识别不仅为用户的回忆提供更为丰富的信息，也让用户能从他人的数据中吸取更多的知识。线路仅仅给用户指明了前行的道路，却并没有告知用户该采用什么方式前往。很多时候开车并不是最佳选择，而有些时候步行也是不可行的。同时，线路之间也可根据不同的交通方式得以区分。显然，当一个自行车爱好者希望查找一条富有挑战性的骑行线路时，我们没有理由要给他推荐一条驾车路线。但是，如果我们不能区分数据库中的线路，那些步行和公交线路将会成为一种误导和噪音。此外，交通方式的识别也是交通状况分析和预测的基础。如果步行的数据被用来预测交通状况，得到的结论将是道路异常拥堵。

在 GeoLife 中，我们首先查找用户的步行路段，然后用这些步行路段对用户的数据进行了分割。分割的规则是来源于日常生活中的规律：在典型的场景下，人们在变换一种交通方式之前必需采用步行，即便步行的长度很短。比如，从公交车下车到乘坐另一辆出租车的过程中，用户必须通过走路来过渡。此后，我们从切分后的轨迹段提取特征，并利用这些特征结合监督学习的方法来训练一个分类模型。然后，给定其他未知交通方式的轨迹数据，我们采用同样的方法进行数据切分和特征提取，并利用训练好的分类模型来识别它的交通方式。基于 60 个用户近 10 个月的数据集，我们得到了 75%以上的预测正确率[9]。

理解用户地理查询日志

近年来，地理信息搜索引擎[11, 12, 13]吸引了互联网上越来越多的眼球。人们经常使用它们来决定行车路线，寻找餐馆以及制定出行计划。随着时间的积累，这些搜索引擎保存了

大量的包含地理信息的用户搜索日志。尽管针对搜索日志挖掘的研究已经开展了多年，但是如何利用这些地理搜索日志在文献中还鲜被提及。

通常来说，一条地理搜索查询包含两部分：1)包含一个以上文本词汇的查询词，2)用户指定的和查询词相联系的地理区域，这里我们称其为搜索位置。需要注意的是，搜索位置可能和用户所处的位置是完全不同的。举例来说，对于一条地理搜索查询”Seattle Pizza”，这里的查询词是”Pizza”，而搜索位置是”Seattle”，搜索用户则可能在全世界的任何一个地方，并不一定要在西雅图。

地理搜索日志的挖掘

在搜索日志挖掘领域，研究人员已经使用了大量不同的算法和技术，例如统计分析[14]，关联规则挖掘[15]，查询词聚类和分类[16]等。在挖掘地理搜索日志时，我们需要将新增的地理信息考虑在内。在本文中，我们将介绍在挖掘语义相关查询词方面 的工作。和传统的工作相比，我们更关心通过地理位置联系的查询词，或称为位置共现模式 [17, 18, 19]。

基于 Monte Carlo 模拟的交叉 K 函数[17]可以被用来测试两类空间对象的共现程度。然而，Monte Carlo 模拟的开销非常大，并不实用。Morimoto[19]第一个定义了在空间数据库中寻找经常相邻的类别集合，即位置共现模式。他们使用支持度(support)，即模式的实例数目来衡量一个模式的显著程度。Shekhar 和 Huang [18] 提出了一个基于 Apriori 算法的挖掘位置共现模式的方法。在[18]中，作者定义了参与度指标(participation index)，一个和简单的相邻对象计数 [19]相比具有更多统计意义的方法，以衡量位置共现模式的显著程度。在这些研究中，位置共现模式的挖掘通常用于挖掘生物和医学数据中的模式，目前还少见将它用于搜索日志的挖掘。

在搜索日志中，位置共现查询词模式的一个例子是{“Children's Museum”, “Experience Music Project”}。这两个查询词的搜索位置在西雅图市是非常接近的。实际上，它们是西雅图市区的两个博物馆，相邻只有几百米。另一个例子是{“shopping mall”, “parking”}。这个模式显示了”shopping mall”和”parking”经常会在相近的位置被搜索。

简短的说，一个位置共现查询词模式包含了一组经常有着相近的搜索位置的查询词。在图 2 中，不同的符号代表着不同的查询词。对于同一个符号，不同的坐标代表一个查询词不同的搜索位置。在图中，查询词 {+, ♦}和 {Δ, O} 是两个位置共现查询词模式，因为每一对查询词都经常在相近的位置被搜索。

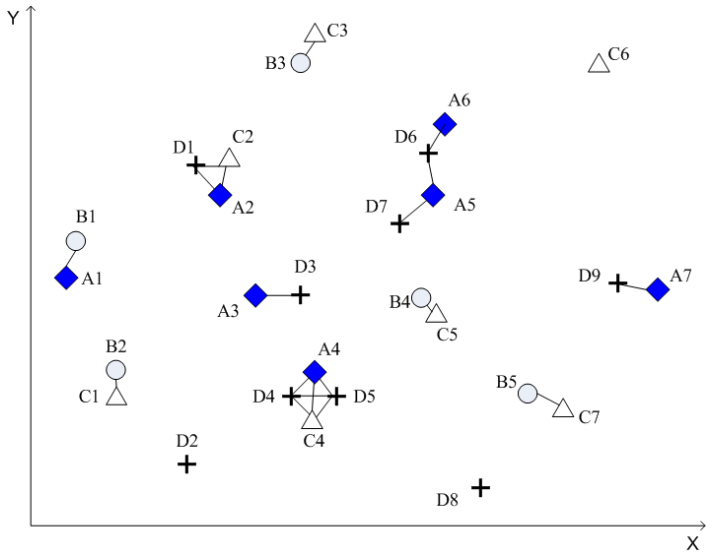


图 2.位置共现模式挖掘

位置共现查询词模式挖掘的应用

挖掘位置共现查询词模式有着广泛的应用，包括查询词推荐，地点推荐，位置相关广告等。

基于从日志中挖掘出的模式，”Sheraton hotel”和”Hilton hotel”是针对查询词”hotel”的很好建议。另外，我们还可以在不同地点为相同的查询词给出不同的查询词建议。例如，我们可以在用户在拉斯维加斯搜索”hotel”时建议”MGM Grand”，因为{“hotel”, “MGM Grand”}是一个出现在拉斯维加斯的位置共现模式。作为比较，我们可以在人们搜索夏威夷旅馆时建议”Ilikai hotel”，因为{“hotel”, “Ilikai hotel”}是一个在夏威夷出现的位置共现查询词模式。

地点推荐可以为用户提供额外的本地信息，以方便他们游览一个不熟悉的城市。例如，当用户在纽约搜索”Time Square” 时，我们可以显示一些在纽约附近的经常被搜索的热门查询词，例如”Twin Tower”，”Ground Zero”和”World Trade Center”。这些热门的位置共现查询词是在搜索位置附近有名的地标。用这种方法呈现位置共现查询词可以让用户依照他们的兴趣浏览目的地附近的景点。

位置共现查询词模式还可以帮助广告商更好的理解本地搜索用户的兴趣。例如，一个查询模式{“car repair”, “Toyota”}显示这个区域的用户可能对 Toyota 汽车的修理服务更加感兴趣。

位置共现查询词模式挖掘算法

在[20]中，我们比较了现有算法[18]和我们提出的一种基于网格划分的方法。现有算法能够发现在整个空间上位置共现的模式。然而，有一部分查询词，尽管不是在整个空间上位置共现，它们在一些特定的区域有着共现的关系。让我们回到{“Children's Museum”, “Experience Music Project”}这个例子。这对查询词实际上并不能为现有算法所发现，这是因为在美国有许多儿童博物馆(Children's Museum)，但是 Experience Music Project 却只有一

个，这样，从全局来看，这个模式并不显著。相似的例子还有{“hotel”, “MGM Grand”}和{“hotel”, “Ilikai hotel”}。 在我们看来，尽管这些查询词不是全局位置共现的，它们在应用中也有着重要的意义。

许多全局的模式实际上是众所周知的知识，而局部模式对人们才更加有趣，特别是对于对本地不太了解的人们。为了解决这个问题，我们提出了一个基于网格划分的模式发现方法[20]。基于网格划分的方法可以发现区域性的位置共现查询词模式，并可以计算模式的局部性，将模式分成局部和全局两类。

直观的来看，局部模式与特定的区域相关。他们可以刻划局部的现象但不能应用到其它地区。另一方面，全局模式则经常是在很多地方出现的现象，例如{“shopping mall”, ”parking”}是一个全局模式，而{“Disney Land”, “EPCOT”'}是一个局部模式。人们通常认为在一个区域发现的模式就是局部模式。然而，他们也可能是全局模式。例如，在很多区域{“shopping mall”, “parking”}都可以被发现为一个位置共现的模式，它实际上是一个全局模式。 在我们的算法中，我们使用模式在不同区域实例数目的熵来表示其局部性。

在我们的用户实验中，用户对使用现有方法找出的模式质量给出了 0.76 的平均分(取值范围是 0 到 1)，而对使用基于网格的方法找出的模式打了 0.9 的平均分。在用户的评价中，使用现有方法找到的模式中的 64%，和基于网格的方法找到的模式中的 94%被标为局部模式。除了高质量的特点，基于网格的方法还可以比现有方法发现更多的模式。根据这些结果，我们得出结论，基于网格的方法要比现有方法更加高效，不管是从模式数量，局部模式的比例和模式质量的角度来看都是一样。

结语

在本文中，我们介绍了最近在从地理角度理解用户行为的工作。特别的，我们介绍了 GPS 轨迹的交通方式分类和位置共现查询词模式挖掘问题。

在未来的工作中，我们还将探索更多类型的数据和用户兴趣。就像所有其它试图理解用户行为的系统一样，数据的隐私性是一个非常重要的问题。在实用系统中，我们必须仔细的设计恰当的机制以保护数据并确信数据是以正确的方式共享给了正确的人。

参考文献

1. Walk Jog Run. http://www.walkjogrun.net/
2. Mountain Bike. http://www.mtb-routes.co.uk/northyorkmoors/default.aspx
3. SportsDo. http://sportsdo.net/Activity/ActivityBlog.aspx
4. D. Ashbrook and T. Starner. Using GPS to Learn Significant Locations and Predict Movement Across Multiple Users. Personal and Ubiquitous Computing 7(5), 2003.

5. J. Krumm and E. Horvitz. Predestination: Inferring Destinations from Partial Trajectories. In Proceeding of UBICOMP'06, California USA, September 2006.

6. D. J. Patterson, L. Liao, and D. Fox. Inferring High-Level Behavior from Low-Level Sensors. In Proc. of the Fifth UBICOMP, 2003.

7. L. Liao, D. J. Patterson, D. Fox, and H. Kautz. Building Personal Maps from GPS Data. IJCAI MOO05, 2005.

8. L. Liao, D. J. Patterson, D. Fox, and H. Kautz. Learning and Inferring Transportation Routines. In Proceedings of the National Conference on Artificial Intelligence, 2004.

9. Y. Zheng, L. Liu, L. Wang, and X. Xie. Learning Transportation Mode from Raw GPS Data for Geographic Applications on the Web. 17th International World Wide Web Conference (WWW 2008), Beijing, China, Apr. 2008.

10. L. Wang, Y. Zheng, X. Xie, and W.-Y. Ma, A Flexible Spatio-Temporal Indexing Scheme for Large-Scale GPS Track Retrieval, the 9th International Conference on Mobile Data Management (MDM 2008), Beijing, China, Apr. 2008

11. Live Local Search. http://local.live.com

12. Google Maps. http://maps.google.com

13. Yahoo! Local. http://local.yahoo.com

14. A. Ntoulas, J. Cho, and C. Olston. What's New on the Web?: the Evolution of the Web from a Search Engine Perspective. Proc. of WWW 2003, pages 1-12, 2004.

15. F. Facca and P. Lanzi. Mining Interesting Knowledge from Weblogs: A Survey. Data and Knowledge Engineering, 53(3):225-241, 2005.

16. D. Shen, J. Sun, Q. Yang, and Z. Chen. Building Bridges for Web Query Classification. Proc. of SIGIR, pages 131-138, 2006.

17. N. Cressie. Statistics for Spatial Data. 1991.

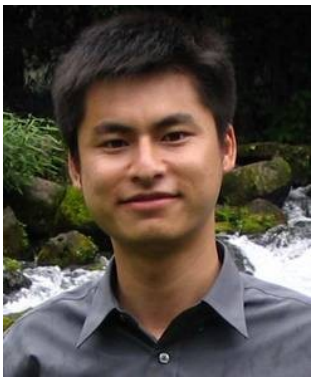
18. Y. Huang, S. Shekhar, and H. Xiong. Discovering Colocation Patterns From Spatial Data Sets: A General Approach. IEEE TKDE, 16(12):1472-1485, 2004.

19. Y. Morimoto. Mining Frequent Neighboring Class Sets in Spatial Databases. In Proc. SIGKDD '01, pages 353-358, New York, NY, USA, 2001.

20. X. Xiao, L. Wang, X. Xie, and Q. Luo. Discovering Colocated Queries in Geographic Search Logs. First International Workshop on Location and the Web (LocWeb 2008), Beijing, China, Apr. 2008.

作者介绍

谢幸博士于２００１年7月加入微软亚洲研究院，现任互联网搜索与挖掘组研究员。他分别于１９９６年和２００１年在中国科技大学获得计算机软件专业学士和博士学位。目前，他主要在移动互联网搜索，基于位置的搜索和移动多媒体应用等方面展开研究。近年来他在国际会议和学术期刊上发表了５０余篇学术论文。他是 ACM 和 IEEE 会员，并多次担任 WWW，CIKM， MDM 和 PCM 等重要国际会议程序委员会委员。



郑宇博士于 2006 年加入微软亚洲研究院，现任互联网搜索与挖掘组副研究员。 他分别于 2001 和 2003 年在西南交通大学获得电机工程专业的学士和硕士学位，并于 2006 年在西南交通大学获得通信与信息工程的博士学位。目前他主要从事基于地理位置的搜索和服务，以及时空数据挖掘和索引等方向的研究工作；近年来他在国内外学术会议和期刊上发表论文 30 余篇。他是 ACM 和 IEEE 会员，并在多个 IEEE/ACM 国际会议中担任程序委员会委员及主席。