Hans Journal of Data Mining 数据挖掘, 2012, 2, 38-42
http://dx.doi.org/10.12677/hjdm.2012.24008 Published Online October 2012 (http://www.hanspub.org/journal/hjdm.html)

Cell Phone User Daily Mobility Pattern Analysis Based on Spectrum Clustering Method

Tao Huang[1], Chen Zhou[2], Benxiong Huang[2], Lai Tu[2]

[1]Wuhan Hongxin Communication Technology Co., Ltd., Wuhan
[2]The Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan
Email: tulai.net@gmail.com

Received: Aug. 9th, 2012; revised: Aug. 22nd, 2012; accepted: Sep. 6th, 2012

Abstract: Along with the development of telecommunication industry and the popularization of mobile phones, cell phones make records of human social behavior data including the call volume, calling patterns, and the location of the cellular phones of their subscribers. How to reveals the rules of human movement behavior based on those data, to make the mobility behavior prediction, has become a rising issue. This article extract characteristics of user mobility and find several kinds of their daily paths though call detail records using spectrum clustering method. The regularity of the same kind of the daily path with different time and special information has been analyzed based on statistic method.

Keywords: Mobility Pattern; Spectrum Clustering; Call Detail Records

基于谱聚类的手机用户日移动行为分析

黄　涛[1]，周　晨[2]，黄本雄[2]，涂　来[2]

[1]武汉虹信通信技术有限责任公司，武汉
[2]华中科技大学电子与信息工程系，武汉
Email: tulai.net@gmail.com

收稿日期：2012 年 8 月 9 日；修回日期：2012 年 8 月 22 日；录用日期：2012 年 9 月 6 日

摘　要：随着通信业的发展和手机的普及，手机记录了大量的人类社交行为数据，其中包括了每个用户每次通话行为以及当时通话的地理位置。如何通过这些数据揭示出人类移动行为的内在规律，从而找到用户的移动特性做出相应的移动行为预测，成为了一个重要的课题。本文通过谱聚类方法，分析手机通话数据，通过提取特征，建立日行为路径相似度的模型，对一个典型用户的日移动行为进行同质归并处理，从而找出以天为单位相同的移动路径。并从星期和活动地域的角度，针对聚类结果中不同簇的日移动路径分别进行了统计分析。

关键词：移动行为模式；谱聚类；用户通话清单

1. 引言

随着信息化时代的发展，人们的生活越来越离不开手机，手机通信数据里隐藏着大量用户移动的规律，对这些规律的研究工作不但能够提高通信运行商的服务质量、优化网络，而且有助于揭示人类移动行为的动态特性[1]。

目前，越来越多的研究者开始将目光投向通过手机的通信数据，并致力于从中找出用户行为的规律，但研究仍处于起步阶段[2]。文献[2]通过用户移动行为模式的鉴别和分析，认为移动行为模式相近的两个陌生用户会以更高的概率成为朋友。文献[1]通过手机通信数据，统计了其同一位置同一时刻一星期为单位的通话次数的分布。文献[3]通过手机通信数据，对移动

*资助信息：本文受新一代宽带无线移动通信网国家科技重大专项(2010ZX03001-001-02)“TD-SCDMA 增强型网络优化工具研发”资助。

行为模式与一种传染性疾病的比较，找出了两者之间的关联。

在研究手机的通信数据的时候，用户隐私是一个敏感的话题，但正如文献[1]中指出，研究人员分析的是加密后的匿名数据，目的仅是深入的研究人类的移动行为规律及其特征，能够有效的克服传统研究中人类行为数据不易采集的缺陷。

在人们开始试图利用手机的通信数据研究人类移动行为之前，移动行为轨迹是一个热点。作为人类移动行为采集工具的有 GPS、移动无线网络以及蓝牙设备，研究致力于找出个人或是群体的移动规律、轨迹以及交往规律[4]，其他与移动规律有关的课题还包括了动物移动特性[5]、车辆移动轨迹特征[6]、疾病传播特性[7]等，其目的也是通过移动行为测量数据找出规律。

与一般的移动行为数据稍稍不同的是，手机通信数据并没有在时间上均匀的提取用户的位置信息，而且其对地理位置的记录并不精确，仅知道通话事件发生时手机用户大概在哪个区域，而且其区域的大小都不相同[2]。所以，对于这类特定的数据，需要使用不同的方法去挖掘出其中所暗含的行为特性。

本文提出了一种基于手机通信数据的分析方法，以用户一天内经过地点的数据集合为一个单元，通过该用户行走方位定义日移动路径相似性，再通过谱聚类方法将其行为分为若干类。本文第二章介绍了选取数据的规则和相似性的定义；第三章介绍了本方法中选用的谱聚类算法的具体处理过程，第四章中展示了聚类结果以及一些针对性的统计分析，最后一章中得出了本文的结论以及后期工作。

2. 用户移动模型

2.1. 日移动行为指标

已有手机用户 u 连续的 M 天的手机通信数据，包括每次通话事件发生时的日期、时刻、服务基站标识。每个基站都有一定的覆盖范围，不同基站的覆盖范围是不同的。如果手机用户进入该基站的覆盖范围，并打出或接听电话，则该基站为其服务基站。提供此次服务的基站标识被记录在手机通信数据中。

如果合并相同的服务基站，则可以从手机用户 u 第 t 天的手机通信数据中，得到这一天为他服务过的基站标识集合 B_t。通过基站与其覆盖范围的对应关系，可以得到该用户这一天去过的地理区域集合 S_t。令 n_t 为第 t 天基站标识集合 B_t 中不同基站标识的个数，则有 $n_t = n(B_t) = n(S_t)$。

2.2. 移动行为相似度

手机用户 u 第 i 天与第 j 天移动行为的相似度

$$s_{ij} = \frac{n_{ij}}{N_{ij}} = \frac{n(S_i \cap S_j)}{n(S_i \cup S_j)} \qquad (1)$$

其中令 $n_{ij} = S_i \cap S_j$ 为第 i 天与第 j 天该用户都去过的地理区域的个数，而 $N_{ij} = S_i \cup S_j$ 为第 i 天与第 j 天该用户去过的地理区域的总个数。s_{ij} 是一个大于等于零小于等于 1 的实数，当两天去过的区域完全相同则为 1，而当其区域完全不同时则其为零。

相似度 s 度量了两天移动行为模式的差异程度，如果某用户周一周二有规律的上下班，而周末会在家里休息或去购物，则可以得到周一与周二的行为相似程度较高，而周一与周末相似程度较低。

3. 社区挖掘方法：谱聚类

3.1. 相似度矩阵

将手机用户 u 第 i 天的移动行为看做一个点，而两天之间的相似程度看成连接两个点的边，则用户 u M 天的通信数据，可以构造成一个图 $G_u = (N, E)$，其中 N 为点集，其有 M 个点，E 为边的集合，是一个 $M \cdot M$ 的对称矩阵，其对角线元素为 1，E 的第 i 行记录了第 i 个点与其他各点(包括它自己)的连接程度，E_{ij} 是一个大于等于零小于等于 1 的实数，零值代表没有连接，1 代表紧密连接。

本文通过社区挖掘方法，以将图 G_u 划分为若干个子图的方式，将用户的日行为分为若干个行为类别。社区挖掘方法大体分为三种，基于优化的算法、启发式和其他[8]，其中谱聚类方法属于基于优化的算法，它将求解最小截问题转化为求解带约束的二次型优化问题：$\min\left\{\left(\boldsymbol{X}^T \boldsymbol{L}\boldsymbol{X}\right)/\left(\mathrm{X}^T \boldsymbol{X}\right)\right\}$ 的方法，其中，向量 $\boldsymbol{X}$ 表示网络划分，$\boldsymbol{L}$ 表示对称半正定矩阵，为拉普拉斯矩阵的不同变体，它可以是 $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ 也可以是 $\boldsymbol{L} = \boldsymbol{D}^{-1/2}(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{D}^{-1/2}$ 或是其他形式[8]，其中矩阵 $\boldsymbol{D}$ 为矩阵 $\boldsymbol{W}$ 的度矩阵，将边矩阵 $\boldsymbol{E}$ 中不为零的数值置

为 1 得到矩阵 $\boldsymbol{W}$，本文采用拉普拉斯矩阵 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$。在聚类方法中，与 k 均值聚类不同，谱聚类不易陷入局部最优值中[8]，故本文采用谱聚类分析，其较为适合分析本文中提出的移动模型的方法。

3.2. 谱聚类原理与算法

文献[9]中表述 2 中提出，矩阵 $\boldsymbol{L}$ 的 k 重特征值为零的特征向量等价于图 $\boldsymbol{G}$ 的 k 个划分，且特征值为零的特征空间被这 k 个特征向量划分[9]，即图 $\boldsymbol{G}$ 非联通子图的个数与 $\boldsymbol{L}$ 的零特征值个数相同，且每一个特征值为 0 的特征向量对应图 $\boldsymbol{G}$ 的一个非联通子图一个社区划分。

其具体步骤为[9]:

Step1 构建表示样本集的相似度矩阵 $\boldsymbol{W}$;

Step2 通过计算 W 拉普拉斯矩阵 $\boldsymbol{L}$ 的任意前 k 个零特征值(小于 $\varepsilon<1*10^{-13}$ 的特征值)与其对应的特征向量,构建特征向量空间;

Step3 利用 k 均值或其它经典聚类算法对特征向量空间中的特征向量进行聚类。

4. 分析结果

4.1. 用户数据初始化

在进行谱聚类之前，要将相似度矩阵 $\boldsymbol{E}$ 所示进行三个步骤的初始化处理，使之满足运算条件。这三个步骤分别是去零行、去孤点和去弱连接。

相似度矩阵 $\boldsymbol{E}$ 某一行对角线元素不为零，这是因为其对应的一天没有任何的通话行为，这一类点需要被清洗掉，否则无法计算。

相似度矩阵 $\boldsymbol{E}$ 某一行除了对角线元素不为零其余全部为零是因为其对应的一天的移动行为不与任何一天相似，在其构造的网络拓扑图中，该点为孤岛。该用户在这一天通话并不充分或当日移动行为异常，无法代表其该天的移动行为轨迹，而且如果保留这类点，会降低聚类运算的灵敏度。

处理后的相似度矩阵 $\boldsymbol{E}$ 如图 1 所示，其中 35 是日期的个数。

第三步需要将相似度矩阵 $\boldsymbol{E}$ 中小于某个阈值的弱连接边去除，因为两个点虽然连接但是连接强度很弱，很有可能是由于一些行为噪声造成的干扰，适当的裁剪一些弱连接的边，会使得社区的主要结构更加的明显。

令裁剪阈值为 Thr，裁剪后的相似度矩阵为 $\boldsymbol{E}_0$ 对应的零特征值的个数为 n_0，其聚类质量度量为 $P=\alpha/SSE+(1-\alpha)\cdot SSB$，其中的簇内凝聚半径为 $SSE=\sum_{i=1}^{K}\sum_{x\in c_i} dist(c_i,x)^2$，簇间分离度 $SSB=\sum_{i=1}^{k}\sum_{j=1}^{k}(c_{i,}\,c_j)^2$，其中 x 为每一天的行为点，c_i 为聚类结果中的第 i 簇的中心点，k 为总的簇个数。当 P 越小时，即簇内凝聚半径越小同时簇间分离度越大时，簇分割质量好。

令 $a=0.5,\ k=3$。裁剪阈值 Thr 与零特征值的个数 n_0 对应关系如图 2 中黑色实心点所示，其中横轴为裁剪阈值 Thr 纵轴为零特征值的个数 n_0。可以看到随

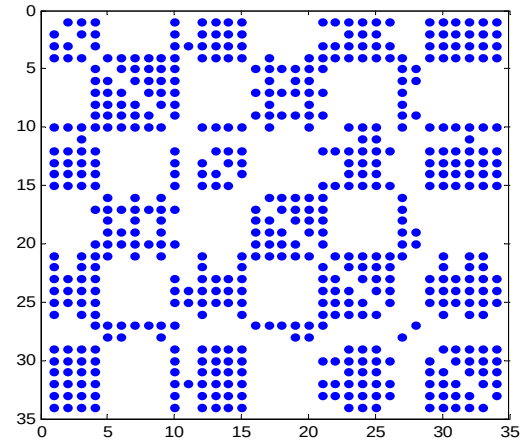

Figure 1. Improved matrix of similarity E

图 1. 处理后的相似度矩阵 E

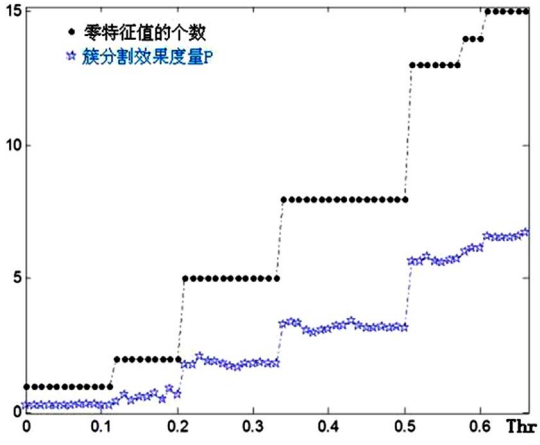

Figure 2. The relationship between clustering partition and Thr and n_0

图 2. 裁剪阈值 Thr 与零特征值的个数 n_0 和簇分割效果度量的对应关系

着裁剪阈值 Thr 的增加其对应的零特征值的个数也呈阶跃式增加。裁剪阈值 Thr 与簇分割效果度量 P 对应关系如图 2 中蓝色五角星点所示，其中横轴为裁剪阈值 Thr 纵轴为簇分割效果度量 P，可以看到 P 也随着裁剪阈值 Thr 的增加而递增，而当仅有一个零特征值时，P 最小。由图可知，$0 < Thr < 0.06$ 是最优的，本文取裁剪阈值 $Thr = 0.03$。

4.2. 谱聚类结果

谱聚类分析后的相似矩阵每列对应类别如图 3 所示，谱聚类分析后的相似度矩阵如图 4 所示，其中横轴 35 为天数，纵轴为簇的类别。

图 3 中横轴代表有通话行为发生的日期，纵轴中 1~3 本别代表了谱聚类的 1~3 类移动行为，可以看到第一类活动行为较多，第三类较少。也可以看出该用户的移动行为较为规律，行动切换的周期性较强，一般会以类 1 行为活动 3~5 天，接着按照类 3 强，一般会以类 1 行为活动 3~5 天，接着按照类 3 行为模式活跃 1 天，再以类 2 移动模式活动 2~5 天。图 4 为谱聚类分析后的相似度矩阵，从左上角到右下角分别为第 1~3 类行为，聚类结果其簇间干扰较少，簇内也较为紧凑。其中类 3 既与类 1 相似又与类 2 相似，在两这类行为间起到过渡作用。

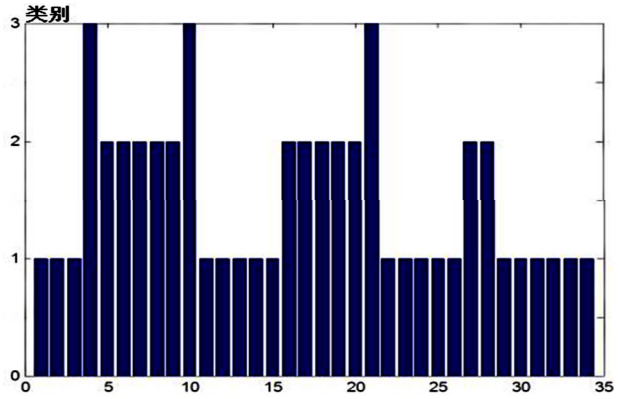

Figure 3. Every corresponding column of the matrix of similarity
图 3. 相似矩阵每列对应类别

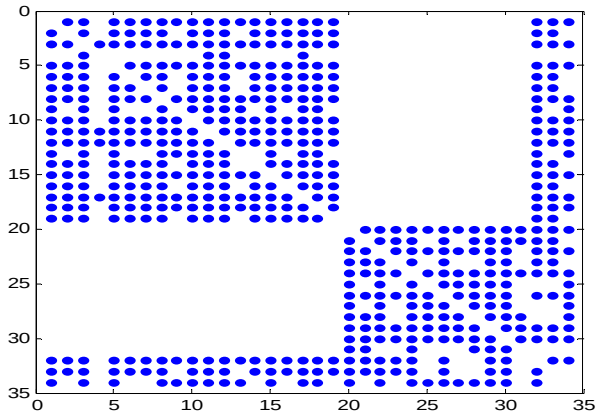

Figure 4. The matrix of similarity after spectral clustering process
图 4. 谱聚类分析后的相似度矩阵

4.3. 每类行为的活动范围统计

统计每类日移动行为所对应的活动范围 $S^i = \bigcup_{t \in C_i} S_t$，其中 S_t 为该用户第 t 天曾去过的地理区域集合，C_i 为第 i 类的日期集合。其 M 天的总活动范围 $\tilde{S} = \{S^1, S^2, S^3\}$ 如图 5 所示。

图中红色点为 S^1 即移动行为类 1 所对应的活动范围，蓝色点为 S^2 即移动行为类 2 所对应的活动范围，黑色点为 S^3 即移动行为类 3 所对应的活动范围。由先验知识可以得知，蓝色点与红色点属于两个不同的城区，其中一个为首要城市一个为二级城市，而黑色点属于郊区。该用户会以天为单位，周期的穿越郊区往返与这两个不同的城区之间。

4.4. 星期统计

以星期为条件，将 M 天的行为数据分为两类，统计周一至周五中三类行为所占的比重，以及周六周日三类移动行为的比例，如下图 6 所示。

由上图可以看出，该用户在工作日类 1 的行为占

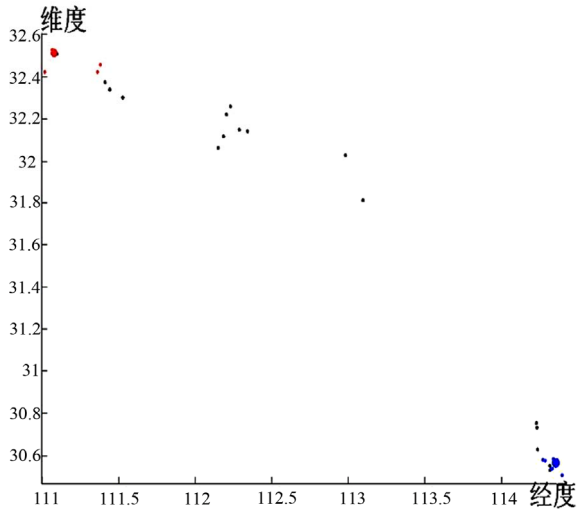

Figure 5. $S = \{S^1, S^2, S^3\}$
图 5. M **天的总活动范围** $S = \{S^1, S^2, S^3\}$

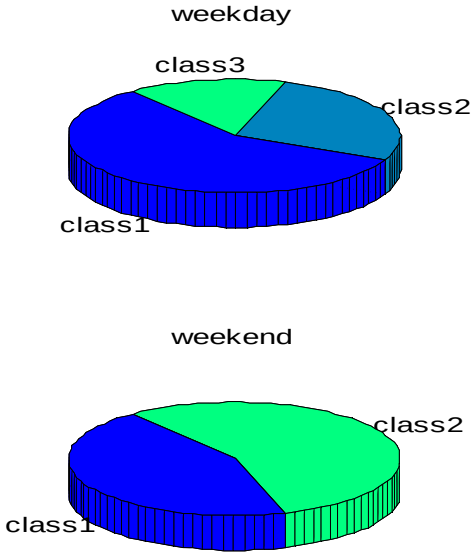

Figure 6. Three behaviors analysis during weekday or weekend
图 6. 周一至周五(weekday)及周六周日(weekend)中三类行为所

大多数，说明其在周一至周五经常在首要城市中活动，其行为类 3 也集中在工作日，说明其时常在工作日在两地间穿梭。而周末时，该用户类 2 的占大多数，说明其在周末会在次要城市中活动，其没有类 3 的行为，说明他很少在周末两地奔波。

5. 结论

本文从手机用户 M 天的通话记录中，提取每天的移动特征，将一天的移动行为看成一个点，构造一个网络拓扑图。通过谱聚类方法，将日移动模式分割成几个不同的类型，并统计了每一种日行为类型的主要活跃地带，以及工作日与周末的因素对该用户该日的移动行为模式的影响程度。

本文选择的典型事例用户的通话行为较多，其地理活动范围较大，所以其日行为的地理切换界限较为明显，其工作日与周末的因素对其移动有一定的影响，而且其在地域间往返的存在周期，其规律性较为明显。

后续工作有几个方向，首先地理活动范围较小而且其基站分布较为分散的情况下，如何去识别用户的日移动行为。因为当地理活动范围较小时，以不同目的的移动行为地域间没有明显的界限；而当基站分布较为分散时，大多数基站都只提供过一次服务，任两天的移动相似性很低，无法准确的找出其不同的移动路径，不适合使用本文提出的移动行为模型。所以，如何建立较之本文更为复杂的行为模型，能够更加准确的更有容错性的找出用户几种不同的常用路径，是有待解决的课题。

第二，对于城镇手机用户，其上班期间与下班之后会有不同的移动行为，如果单以日期为单位鉴别其路径，可能会产生一些混乱无法解释的聚类结果，的是否可以结合时段的分类，进一步的找出用户的移动行为规律，使其预测工作更加的准确。

第三，如何建立一种架构，将有不同的移动路径、不同的移动模式的用户进行区分和统计。针对不同的用户提出不同的分析方法与处理参数，当增加新用户的时候，能够更快速准确的对其进行定位与分析。

参考文献 (References)

[1] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey and A. L. Barabási. Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical, 2008, 41: Article ID: 224015.

[2] D. Wang, D. Pedreschi, C. Song, F. Giannotti and A. L. Barabási. Human mobility, social ties, and link prediction. 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'11), 2011.

[3] J. Tatem, Y. Qiu, D. L. Smith, O. Sabot, A. S. Ali and B. Moonen. The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents. Malaria Journal, 2009, 8(1): 287.

[4] M. Musolesi, C. Mascolo. Mobility models for systems evaluation: A survey. Middleware for Network Eccentric and Mobile Applications, 2009: 43-62.

[5] R. M. Fewster, C. Southwell, D. L. Borchers, S. T. Buckland and A. R. Pople. The influence of animal mobility on the assumption of uniform distances in aerial line-transect surveys. Wildlife Research, 2008, 35(4): 275-288.

[6] C. Curtis, T. Perkins. Travel behaviour: A review of recent literature, 2006. http:www.urbanet.curtin.edu.au/local/pdf/ARC_TOD_Working_Paper_3.pdf

[7] C. Martin, P. P. Pastoret, B. Brochier, M. F. Humblet and C. Saegerman. A survey of the transmission of infectious diseases/infections between wild and domestic ungulates in Europe. Veterinary Research, 2011, 42(1): 70.

[8] B. Yang, D. Y. Liu, L. Jiming, D. Jin and H. B. Ma. Complex network clustering methods. Journal of Software, 2009, 20(1): 54-66.

[9] U. Von Luxburg. A tutorial on spectral clustering. Statistics and Computing, 2007, 17(4): 395-416.