GPS 位置历史挖掘和移动轨迹异常检测系统的设计与实现*

熊 妍[1,2]，罗 泽[1†]，阎保平[1]

(1. 中国科学院计算机网络信息中心，北京 100190；2. 中国科学院大学，北京 100049)

摘 要：为了发现候鸟的公共移动模式和异常迁徙行为，通过对候鸟位置历史数据建模，设计与实现了位置历史挖掘和异常检测系统，分析并可视化展示候鸟迁徙过程中经停地和移动序列的分布状况及其重要性。通过计算候鸟移动序列相符度并可视化展示结果，发现候鸟迁徙的公共模式和异常行为。使用青海湖斑头雁的卫星遥测数据验证了这一系统，结果发现了斑头雁的三个公共移动模式和三个异常迁徙行为，验证了该系统的有效性。

关键词：公共模式；异常检测；移动序列相符性；斑头雁

中图分类号：TP39　　**文献标志码**：A　　**文章编号**：1001-3695(2013)09-2704-03

doi:10.3969/j.issn.1001-3695.2013.09.036

Design and implementation of GPS location history data mining and anomalous detection system

XIONG Yan[1,2], LUO Ze[1†], YAN Bao-ping[1]

(1. *Computer Network Information Center, Chinese Academy of Sciences, Beijing* 100190, *China*; 2. *University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract: This paper designed and implemented a system for modeling GPS location history data and anomalous detection. This system could analyze and visualize the distribution of stopovers, movement sequences of moving objects, and the importance of the stopovers. By calculating and visualizing the movement sequences consistency, this system could discover common movement sequences and uncommon movement behaviors. It conducted experiments on the data collected from bar-headed geese captured in the Qinghai Lake region and discover three common movement sequences and three anomalous behaviors respectively which indicate the correctness of our system.

Key words: common movement sequences; anomalous detection; movement sequences consistency; bar-headed geese

近年来随着位置感知技术的发展，使得人们能够以较高的时空分辨率记录位置历史数据。然而，高时空分辨率位置历史数据的急剧增长，对开展数据分析和挖掘提出了新的挑战[1]。因为位置历史数据的分析高度依赖于应用的需求。基于全球定位系统的位置感知设备，通过在特定时间间隔记录并存储移动对象的几何位置(基于 WGS84 大地坐标系统的经、纬度和高度)获得其位置历史信息。但是，位置感知设备无法记录相关的背景信息和应用相关的语义信息，而这些信息的缺失会妨碍对数据进行有效和智能的分析。

当前，无论是针对个人还是针对候鸟的位置历史数据都开展了一些研究工作。然而对于两者的研究，除了研究对象、数据采集频率要求不同外，研究的目的也不一样。候鸟具有相对稳定的迁徙行为，如迁徙周期和迁徙路线；而对人的位置信息的研究主要集中在提供位置感知的商业增值服务，如针对用户移动行为模式的广告投放[2]、根据用户相似性的旅游线路推荐[3]、国土安全防卫等；而迁徙候鸟的位置历史数据分析集中在对候鸟的迁徙行为和模式开展分析[4,5]，研究候鸟的活动范围、迁徙地选择、迁徙路线，为疾病传播、灭绝风险分析、生态研究提供支持[6,7]。

Tang 和 Carneiro 提出的方法都是通过基于点的聚类方法对迁徙候鸟的位置历史数据进行分析。这种方法能够反映总体行为特征(总体的迁徙路线、主要经停地)，但却忽略了个体的行为特征。Muzaffar 逐个分析个体行为特征，工作量大，且不能直观地发现总体行为特征。

本文研究青海湖地区斑头雁(anser indicus，拉丁语)的迁徙行为，发现每只斑头雁的迁徙行为和迁徙路线，以及在迁徙过程中选择的栖息地和经停地。在此基础上，分析其经停地的重要性。通过计算任意两只斑头雁的移动序列相符度，展示斑头雁种群的公共移动模式以及存在的异常行为。

1 系统架构

本文提出位置历史挖掘和异常检测(location history mining and anomalous detection，LHMAN)系统来分析每只鸟的移动序列和候鸟迁徙过程中产生的经停地集合，以及每个经停地的重

收稿日期：2012-12-21；**修回日期**：2013-02-02　　**基金项目**：国家自然科学基金资助项目(90912006)；国家科技基础条件平台建设项目(BSDN2009-18)

作者简介：熊妍(1987-)，女，河南信阳人，硕士，主要研究方向为数据挖掘、机器学习(xiongyan@cnic.cn)；罗泽(1976-)，男(通信作者)，副研究员，博士，主要研究方向为数据挖掘、面向服务架构等；阎保平(1950-)，女，研究员，博士，主要研究方向为 e-Science 应用示范、大规模数据集成和处理.

要性,通过计算候鸟移动序列的相符度评价候鸟移动行为的相似性,并进行可视化展示。系统设计如图 1 所示,分别包含移动数据存储、数据分析和可视化。

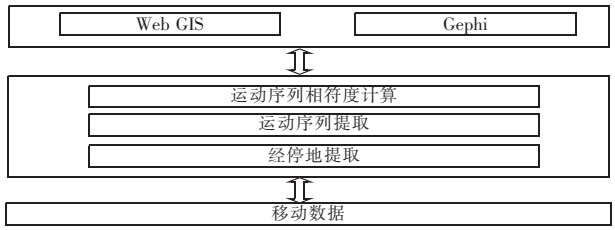

图 1　LHMAN 系统框架

1)移动轨迹数据存储　系统使用关系型数据库存放位置历史数据,并保存数据分析阶段产生的中间文件。

2)移动轨迹历史数据分析　数据分析包括三个部分,当接收到来自用户的请求时,经停地识别模块便会从数据库中查询相应数据并运行经停地提取和生长算法,然后提取移动序列,并通过移动序列相符度模块计算个体之间的相似度。

3)可视化展示　该模块包含两部分,即通过 GIS 可视化展示经停地的分布、候鸟移动路径等相关信息和利用 Gephi 可视化展示候鸟移动序列相符度结果。

2　移动轨迹数据设计

LHMAN 系统后台使用关系型数据库,存储 GPS 位置历史数据以及数据分析模块产生的经停地、移动路径等数据。

2.1　GPS 位置历史数据

本研究的数据来自在青海湖捕获的 29 只承载了卫星遥测设备的斑头雁。其中 14 只是在 2007 年 3 月 25 ~ 31 日被捕获的,其他的斑头雁是在 2008 年 3 月 28 日 ~4 月 3 日被捕获的。每只斑头雁被称重、测量并且绑定一个重 45 g 的太阳能供电便携式遥测设备,包括一个终端传输(PTT:9 NorthStar Science and Technology,LLC,Baltimore,Maryland USA)和一个微波遥测终端(PTT-100,Columbia,Maryland USA),可以同时通过 Argos 卫星和 GPS 接收器进行定位。

经预处理后,位置历史数据格式如图 2 所示。

Animal	Record_type	Datatime	Latitude	Longitude	Lc94
BH07_67586	Arog8	2007-08-23 06:44:40	34.752	98.1830	L1
BH07_67586	GPS	2007-08-23 09:00:00	34.7583	98.15730	LG

图 2　位置历史数据格式

由于每只斑头雁的数据量不一致,为了更好地发现斑头雁的迁徙规律,本文舍去了 2 只数据量不足一个月的斑头雁,并将每只具有超过一年数据的斑头雁按年分段,获得 33 条路径分析。

2.2　经停地数据

经停地提取模块产生的经停地信息包括基本信息和组成经停地轮廓的多边形坐标信息,如图 3 和 4 所示。

id	lat	lng	area	num	times	days
2	90.904	29.546	373 km^2	14	15	1909

图 3　经停地基本信息数据格式

id	lat1	lng1	lat2	lng2	lat3	lng3	lat4	lng4
2	86.834	20.482	86.908	20.521	86.966	20.66	…	…

图 4　经停地轮廓坐标信息

2.3　移动路径数据

移动路径的基本信息包括途径的经停地序号、频度等,如图 5 所示。其中,id 表示移动序列的编号,sp_1、sp_2、sp_3 为移动路径 1 途径的经停地集合,freq 表示遵循此路径鸟的个数。

2.4　移动路径相符度数据

相符度是两条移动路径之间相似度的度量。其格式如图 6 所示。其中 seq_1、seq_2 为移动路径的编号。

id	sp_1	sp_2	sp_3	freq
1	8	6	2	10

图 5　移动路径信息数据格式

id	seq_1	seq_2	consistency
1	1	2	0.8

图 6　相符度数据格式

3　移动轨迹历史数据分析

首先描述在本文中使用的几个概念。

点:候鸟某一时刻在空间中的位置,主要包含时间戳 t、纬度 lat 和经度 lng 字段。

路径:候鸟在某一段时间内的移动轨迹。一条路径是一个按照时间顺序排序的点的集合,记做$\{p_1,p_2,p_3,\cdots,p_n\}$,其中 $p_i.t<p_{i+1}.t,1\leqslant i<n$。

经停地:代表地理空间中的一块区域。经停地具有四个属性,即纬度 lat、经度 lng、到达时间 arv$T$、离开时间 lev$T$。其中,$\text{lat}=\frac{1}{k}\sum_{i=1}^{k}p_i.\text{lat}$, $\text{lng}=\frac{1}{k}\sum_{i=1}^{k}p_i.\text{lng}$, $\text{arv}T=p_i.\text{arv}T$, $\text{lev}T=p_i.\text{lev}T$ $p_i(1\leqslant i\leqslant k)$代表组成此经停地的 k 个点。

3.1　经停地提取

设定一个时间阈值 th 和一个距离阈值 d_t,按照时间顺序逐点扫描每只鸟的轨迹。给定一个点序列$\{p_1,p_2,p_3,p_4,\cdots,p_k\}$,从当前点开始,逐个检查后续点,直到距当前点的聚类超过阈值 d 或者时间超过 th,则这部分点构成一个经停地,然后继续检测剩余点找到更多的经停地,即$\{D_j|\text{Distance}(p_i,p_j)\leqslant d$ 且$|p_k.t-p_i.t\geqslant th,i\leqslant j\leqslant k\}$, 构成的区域为一个经停地。

图 7 中示例了两种类型的经停地。p_2 构成的经停地 sp_1 表示对象长时间地待在同一个地方,如鸟在夜间的栖息地;p_3、p_4、p_5、p_6 构成的经停地 sp_2 表示对象在一个小的空间范围内进行较长时间的小范围活动,如鸟在某个区域停下来觅食。

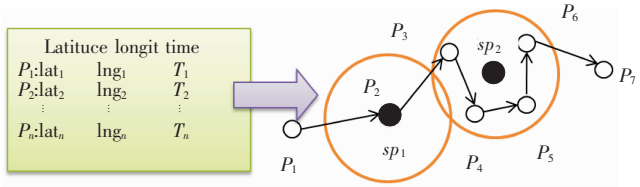

图 7　点、路径、经停地示意图

从数据集中提取出每只候鸟的具体经停地之后,使用经停地生长算法提取候鸟种群的经停地。该算法简单描述如下:如果任意两个经停地在地理上有重叠,便将这两个经停地合并,并赋予最终得到的每个经停地一个唯一标志,用于后序序列提取。

3.2　移动序列提取

每只鸟的移动序列可以表示为一个按照时间排序的经停地集合。经过经停地提取后,每只鸟的移动序列表示为其移动路径包含的经停地所对应的 ID 序列。如果连续多个经停地的 ID 相同,则在结果序列中用其 ID 与次数表示。

在图 8 中,沿着箭头的方向,Bird1 的路径用经停地编号表示为$\langle A\rightarrow B\rightarrow C\rightarrow C\rightarrow D\rightarrow D\rangle$,处理重复经停地后,最终表示为

〈$A(1)\rightarrow B(1)\rightarrow C(2)\rightarrow D(2)$〉。Bird2 的路径表示为〈$A\rightarrow C\rightarrow B\rightarrow C\rightarrow B\rightarrow B$〉,处理重复经停地后,最终表示为〈$A(1)\rightarrow C(1)\rightarrow B(1)\rightarrow C(1)\rightarrow B(2)$〉。括弧中的数字表示在此路径中连续落入此区域的经停地的个数。

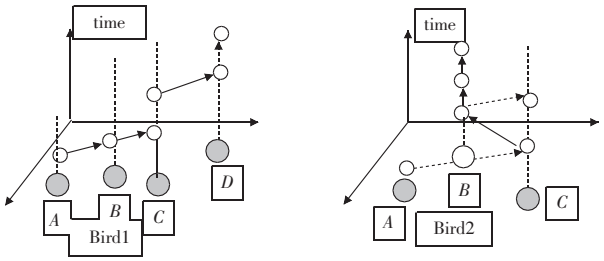

图 8 移动序列提取

3.3 移动序列相符度计算

移动序列提取完成后,计算任意两个移动序列的相符度。首先提取相符序列,然后根据相符度计算方法计算相符度。最终 M 只鸟的相符度结果可表示为一个 $M\times M$ 的矩阵 S,元素 $s[i][j]$ 表示鸟 i 与鸟 j 之间的移动序列相符度。

3.3.1 相符序列提取

相符序列指的是两个移动序列中最长的重合部分。例如图 9 中,鸟 1 的移动序列为〈$A(1)\rightarrow B(1)\rightarrow C(2)\rightarrow D(2)$〉,鸟 2 的移动序列为〈$A(1)\rightarrow C(1)\rightarrow B(1)\rightarrow C(1)\rightarrow B(2)$〉,则两者最长的相符序列为〈$A(1)\rightarrow B(1)\rightarrow C(1)$〉,其中每一个经停地累计出现的次数取两者之中的最小值。

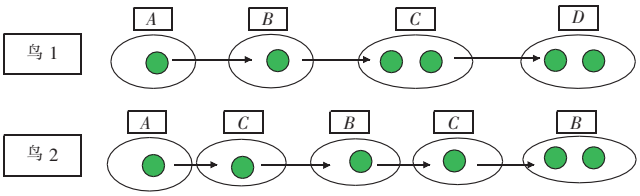

图 9 提取移动序列的相符序列

3.3.2 移动序列相符度计算

最短编辑距离是一种利用动态规划的思想进行序列之间相似度计算的方法。鸟 b_1 与 b_2 之间的相符度计算方法为:

a)利用最短编辑距离算法分别计算鸟 b_1、b_2 的移动序列与两者的相符序列之间的编辑距离,分别记做 ld_1、ld_2。

b)鸟 b_1 与鸟 b_2 的相符度为 $s[1,2]=1-ld_1/b_1.\text{sequence}$;而鸟 b_2 与鸟 b_1 的相符度为 $s[2,1]=1-ld_2/b_2.\text{sequence}$。

4 可视化展示

可视化展示模块用网络可视化工具 Gephi 和 Web GIS 展示经停地、移动路径的详细信息以及候鸟的移动序列相符度结果。

4.1 经停地展示

图 10 展示了 33 只青海湖斑头雁在迁徙过程中途径的 10 个经停地分布状况,点击可查看该经停地的经度、纬度、面积、途径该经停地的统计信息。它们按照纬度从低到高依次编号为 1,2,3,…,9,10。其中,1 号经停地为印度 Bhitarkanika 国家森林公园,2 号经停地为西藏拉萨,6 号经停地为扎陵湖、鄂陵湖,8 号经停地为青海湖,10 号经停地为蒙古后杭爱省。

4.2 移动路径展示

图 11 为青海湖斑头雁的 33 个移动序列图。其中,不同的序列用不同颜色表示。线的宽度与遵循此移动序列的鸟的数量成正比。点击线段可以查看该移动序列途径的经停地。直观来看,可以发现遵循路线青海湖—扎陵湖、鄂陵湖—拉萨的青海湖斑头雁最多。

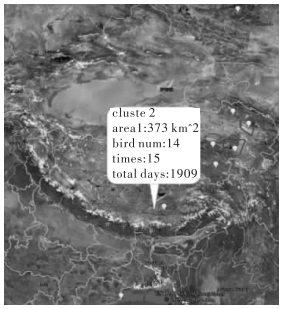

图10 青海湖斑头雁迁徙经停地分布

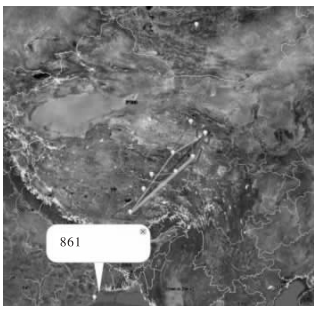

图11 青海湖斑头雁移动序列图

4.3 相符度结果展示

相符度计算结果如图 12 所示。图中每一个节点代表一只鸟,边代表两只鸟之间的移动序列相符度。为了发现候鸟移动的公共模式和异常行为,系统只显示相符度为 1 的鸟,即如果两只鸟的移动序列之间是包含与被包含的关系;否则就在图中表现为孤立的点。

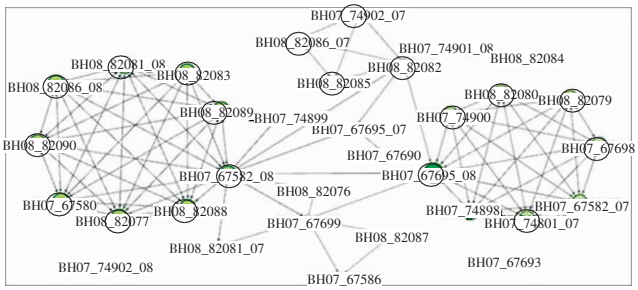

图 12 移动序列相符度结果

由图中可知,33 条的移动序列彼此之间呈现一定的共性。24 个具有完整的全年数据,其中 23 个的迁徙方向是从青海湖向西南方向迁徙。58%(14/24)春夏两季栖息在青海湖,途径扎陵湖、鄂陵湖和格尔木市,向西南方向迁徙至西藏拉萨越冬;另外还有 8%(2/24)遵循这一迁徙路线,只不过在西藏那曲越冬;13%(3/24)在青海度过繁殖期,但是会迁徙至扎陵湖、鄂陵湖越冬;剩下的 20%(5/24)不遵循任何一条公共模式。由此可以发现,大多数青海湖斑头雁在青海湖结束繁殖期之后,会在西藏的东部、北部越冬。公共移动序列是在青海繁殖,途径扎陵湖、鄂陵湖到达拉萨越冬。这一结果与文献[4]中的观察和实验一致。

图中有 5 只鸟是独立的。检查它们的移动序列,发现它们不遵循公共移动序列。其中有 2 只鸟(NO. BH07_74902_08,NO. BH07_82084)在青海湖西北方向的哈拉湖繁殖;1 只鸟(NO. BH07_67693)在 2007 年的繁殖期没有像别的鸟一样在青海湖繁殖,而是在蒙古后杭爱省度过繁殖期;1 只鸟(NO. BH08_82081)2008 年在青海湖栖息,但是飞至印度的 Bhitarkanika 国家森林公园越冬;1 只鸟(NO. BH07_74901)2007 年在青海湖繁殖,在拉萨越冬,但是在 2008 年的夏天,它却栖息于青海玉树,并在那曲越冬。

5 结束语

本研究的创新之处在于:a)提出了一种基于位置历史挖掘和异常检测的系统框架,该框架可对候鸟的 (下转第 2710 页)

3.4 基于云计算的解析服务系统[6]

云平台是云计算系统的核心组成部分。它作为云计算服务的基础,管理着数量巨大的底层物理资源,以虚拟化的技术来整合一个或者多个数据中心的资源,屏蔽不同底层设备的差异性,统一分配和调度计算资源、存储资源与网络资源,以一种透明的方式向用户提供包括计算环境、开发平台、软件应用在内的多种服务。在使用者看来,云平台的资源是无限扩展的,用户可以利用各种终端设备通过网络接入到云平台,随时获取、实时使用、按需扩展计算和存储资源。

物联网通过覆盖全球的传感器、RFID 标签技术实时感知海量数据,并能通过汇聚、挖掘与智能处理获取有价值的信息,为不同行业的应用提供智能服务。物联网数据具有海量、多态、动态、关联的特点。全球数以万计的生产厂商,每时每刻都在进行物品的生产,于是在物流运输方面,物品的动态信息在变化;商品交易方面,物品所有者也在变化;在人们获取其使用价值时,物品的状态信息也会变化。这些物品的数据需要同步到云上,因此要利用数据中心与计算平台存储、管理物联网的海量数据,并对物品的编码进行解析服务,然后反馈需要的物品信息。

本文所描述的编码方法就是把单品编码与它所对应的信息建立映射关系,如式(1)~(3)所示。单个物品的信息就由 e、M、N 以及物品所有者和物品的信息组成,物品的状态信息随物品所有者动态更新。用户还可以根据自己的物品情况下载合适的数据库资源,进行动态更新,以满足解码的需要。

3.5 新编码体系的特点

在本文提出的编码体系中,以地理坐标确定生产厂商,打破了其他编码体系中国家、地区的限制,建立起三维空间坐标点与全球的厂商单位一一对应,并且能够无限扩展系统。物品类别代码由厂商决定,不再受国际组织制定的各种物品分类标准的束缚,克服了现在分类标准对未来新领域新事物界定不明确的缺点。此编码体系只有三段编码要素,代码简洁,利用云计算技术进行物品代码解析,速度更快。对实体物品的抽象分类,并且每类都采用相同的编码方法,打破了不同行业和不同领域的限制,突破了现存编码体系分类标志不统一,方案不兼容,无法实现跨行业、跨平台、规模化的物联网应用。

4 结束语

本文分析了现存编码体系的利弊和现实生活中物品属性,提出一种新型的统一编码体系和思想,抽象出载体物品和实体物品,并阐述了两者之间的关系,对应现实中四种情况的编码问题,采用同一种新的编码方法,解决了物品的生产和流通阶段的统一编码难题。进一步解释了物品在销售和使用阶段解析信息的更新方法,满足物联网每一个物品情况的实时监控和智能化管理。

参考文献:

[1] 徐珉. 基于 EPC GID-96 矿用物联网标签编码方案研究[J]. 煤炭科学技术,2012,40(5):70-73.

[2] KONG Ning, LI Xiao-dong, YAN Bao-ping. A model supporting any product code standard for the resource addressing in the Internet of things[C]//Proc of the 1st IEEE International Conference on Intelligent Networks and Intelligent Systems. Washington DC: IEEE Computer Society, 2008: 233-288.

[3] 熊世娟. 基于 RFID 的物联网标识兼容模型与兼容机制的研究[D]. 北京:北京邮电大学,2011.

[4] 刘冬冬. 基于 PZP 的物联网信息发现服务的研究[D]. 郑州:郑州大学,2011.

[5] 张铎. 物联网与物品标识系统[J]. 物联网技术,2012,3(1):1-4.

[6] 吴功宜,吴英. 物联网工程导论[M]. 北京:机械工业出版社,2012:249-261.

[7] 宁焕生,徐群玉. 全球物联网发展及中国物联网建设若干思考[J]. 电子学报,2010,38(11):2590-2598.

[8] EPC global Inc. V1.0.1, Object naming service(ONS) standard[S]. US: EPCglobal, 2008.

[9] 柴欣. 物联网信息技术在现代食品物流中的应用[J]. 物流工程与管理,2012,34(8):45-46.

[10] 刘尧,高峰,徐幸莲,等. 基于 RFID/EPC 物联网的猪肉跟踪追溯系统开发[J]. 食品工业科技,2012,33(16):49-52.

[11] 刘学江. 铁路物流物联网体系架构[D]. 成都:西南交通大学,2012.

[12] 刘鹏程. 浅谈物联网与物品编码标准化[J]. 物流技术,2011,30(1):17-24.

(上接第 2706 页)GPS 数据进行有效的分析,在此基础上实现的移动位置历史挖掘和异常检测系统能够分析并可视化展示候鸟迁徙过程中产生的经停地分布、每一个经停地对候鸟迁徙的重要性以及移动序列分布;b)提出了一种移动序列相符度计算方法,可视化展现该计算结果后,可轻松发现常见移动模式和异常行为,为生物学家进行动物监测和分析提供了基础。目前,在该系统所采用的建模方法中,在提取移动序列时只考虑了序列的顺序,没有考虑候鸟栖息在各个经停地的时间顺序,而后者也应该是计算移动序列相符度时应该考虑的因素。在接下来的工作中,会逐步完善该模型。

参考文献:

[1] SMOUSE P E, FOCARDI S, MOORCROFT P R, *et al*. Stochastic modeling of animal movement[J]. Philosophical Transactions of the Royon Society B, 2010, 365(1550): 2201-2211.

[2] LI Quan-nan, ZHENG Yu, XIE Xing, *et al*. Mining user similarity based on location history[C]//Proc of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM Press, 2008.

[3] ZHENG Yu, ZHANG Li-zhu, MA Zheng-xin, *et al*. Recommending friends and locations based on individual location history[J]. ACM Trans on the Web, 2011, 5(1): 1591-1596.

[4] TANG Ming-jie, ZHOU Yuan-chun, LI Jin-yan, *et al*. Exploring the wild birds' migration data for the disease spread study of H5N1: a clustering and association approach[J]. Knowledge and Information Systems, 2011, 27(2): 227-251.

[5] CLAUDIO C, ARDA A, JOSE M, *et al*. Advanced data mining method for discovering regions and trajectories of moving objects: ciconiacicconia scenario[C]//Proc of AGILE. Berlin: Springer, 2008: 201-224.

[6] CUI Peng, HOU Yuan-sheng, XING Zhi, *et al*. Bird migration and risk for H5N1 transmission into Qinghai Lake, China[J]. Vector-Borne and Zoonotic Disease, 2011, 11(2): 567-576.

[7] MUZAFFAR S B, TAKEKAWA J Y, PROSSER D J, *et al*. Seasonal movements and migration of Palla's Gulls Larus ichthyaetus from Qinghai Lake, China[J]. Forktail, 2008, 24(2008): 100-107.