基于用户轨迹的*POI*个性化推荐算法研究

李 坡，华一新，李 响，范林林，冯长强

（信息工程大学 地理空间信息学院，河南 郑州 450001）

摘 要：为了实现兴趣点（*POI*）的个性化推荐，本文针对用户轨迹中的含有大量冗余点的问题，探讨了利用*POI*和公交数据对用户轨迹数据进行压缩的算法。研究了传统的协调过滤推荐算法后，提出一种基于用户轨迹的加权TopN推荐算法（UserTN）。实验结果表明，推荐结果的准确率、召回率和个性化程度都优于传统的协同过滤推荐算法，证明了本文方法的有效性。

关键词：用户轨迹；相似性；数据压缩；*POI*；推荐算法

中图分类号：P228.4 **文献标识码**：A **文章编号**：1672-5867(2016)11-0055-04

Research on *POI* Personalized Recommendation Algorithm Based on Users' Trajectory

LI Po, HUA Yi-xin, LI Xiang, FAN Lin-lin, FENG Chang-qiang

(**Institute of Surveying and Mapping, Information Engineering University, Zhengzhou 450001, China**)

Abstract: In order to realize personalized recommendation of the point of interest (*POI*), aim at contain a lot of redundant points in the users´trajectory; discuss the *POI* and traffic data are used to compress users' trajectory data. After studying collaborative filtering recommendation algorithm, In this paper, a weighted TopN recommendation algorithm based on user trajectory (UserTN) is put forward. The experimental results show that the accuracy, recall rate and individuation degree of recommended results is superior to the traditional collaborative filtering recommendation algorithm, proves the validity of this algorithm.

Key words: users´trajectory; similarity; data compression; *POI*; recommendation algorithm

0 引 言

目前，基于站内用户行为分析的个性化推荐服务已广泛存在。如爱奇艺视频网站在“猜你喜欢”一栏推荐用户可能喜欢的视频；京东商城根据用户的喜好精心为用户个性化推荐其感兴趣的商品；QQ音乐在其“电台”菜单下设置“猜你喜欢”子集向不同用户推荐歌曲。

随着GPS、基站和WI-FI定位等技术的快速发展，很多基于位置服务的软件记录了大量用户站外轨迹数据。利用站外数据对用户行为分析的研究开始出现，如何运用站外轨迹数据对用户的喜好进行判断和预测，并进行个性化推荐，成为相关企业和学者研究的一个热点。

本文通过研究目前传统的协同过滤推荐算法可知，这些算法对基于轨迹向目标用户推荐*POI*的效率和个性化程度都较低。所以，本文对以下几个方面进行改进来提高*POI*推荐的个性化程度。首先，由于GPS轨迹的精准，致使数据量大、轨迹相似性分析难度大，本文引入了*POI*和公交数据，提出了一种用户轨迹简化模型；其次，通过相关研究可知，不同年龄段和性别的人群喜好差别较大，所以，本文依据属性对用户集进行划分；最后，在个性化推荐研究中，本文改进了传统的用户相似算法，并根据相似度确定权重，提出了一种加权TopN推荐算法。

1 基于道路和*POI*的轨迹数据压缩算法

随着经济的发展和技术的成熟，位置服务软件普遍存在于用户手机上。在使用过程中，产生了海量轨迹数据。据统计，如果每30 s采集一次GPS数据，则30人在1 d就可达到1 G的数据。随着时间和用户数的增大，将会产生巨量的轨迹数据，且轨迹中含有大量的冗余点。如此庞大的坐标点如不采取数据压缩技术，利用轨迹信息计算用户的相似性时就会出现耗时长、精度差等问题。

轨迹数据压缩算法主要划分为两种，一是分段线性

收稿日期：2016-04-05

基金项目：国家自然科学基金（06023201）资助

作者简介：李 坡（1991- ），男，河南滑县人，地图制图学与地理信息工程专业硕士研究生，主要研究方向为用户行为分析、数据可视化等。

化用户轨迹数据,其算法形式较为简单,计算复杂度低且耗时段,所以成为最常用的压缩算法;另一种是非线性的轨迹拟合,该算法更接近真实轨迹,但是算法复杂、计算量大。用户轨迹受到道路的限制,所以线性化的表示方法就能够很好地描述现实中用户的运动状态。Top - Down 算法是具有代表性且效率较高的线性化压缩算法。

Top - Down 算法根据首末点确定轨迹的大致方位,然后计算所有系列点在该方向的偏离值,若偏离值大于给定阈值的即认定为轨迹特征点,将所有轨迹特征点存储,完成数据压缩[1]。所以,该算法只适用于有明确的起点和终点的轨迹数据。在对道路约束条件下的轨迹数据压缩时,采用 Top - Down 算法不能反映用户的实际行为。基于该特征,本文提出了基于道路和 *POI* 的轨迹数据压缩算法。主要分为以下步骤:

1)数据叠加。将 *POI*、用户轨迹、道路和公交路线等数据的坐标系统转换为西安 1984 坐标系,并将这些数据进行叠置处理,生成新的地图。

2)指定规则,剔除无用数据。根据用户轨迹,将距离轨迹路线 30 m 以上的 *POI* 点剔除数据集,并保留与用户轨迹有 2 站以上的重合路线的公交数据和相应的道路数据。

3)重构用户轨迹数据。基于上述数据,若轨迹数据与公交路线处于同一道路上,则替换为公交路线,并将始点与终点标注为 *POI* 节点。其余轨迹数据以 *POI* 为节点和地图道路为标准简化轨迹。首先,将道路的中心线作为标准线,计算所有数据点(除去起始点)到标准线的垂直距离,得到所有的距离;其次,得到所有大于阈值 A 的节点,删除其他节点;再次,计算剩余节点与 *POI* 点距离,如果小于阈值 B,则删除最小距离对应节点,否则保留节点,删除 *POI*;最后,连接保留下来的起点、终点、节点和 *POI*,生成新的轨迹。

如图 1 所示,实线为用户初始轨迹,虚线为规则后轨迹,方点为起点和终点,圆点为 *POI* 点,三角形点为轨迹节点。

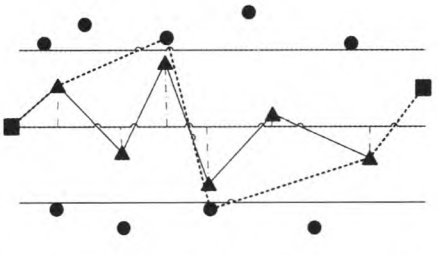

图 1 基于道路和 *POI* 的轨迹数据压缩算法

Fig. 1 Trajectory data compression algorithm based on road and POI

2 基于用户轨迹的 *POI* 推荐算法研究

个性化推荐技术通过研究用户的喜好和兴趣,为用户推荐其所须的各种资源,最初应用于电子商务个性化服务中[2]。而在诸多推荐算法中,协同过滤推荐算法是现今最为成功、应用最广泛的推荐方法[3]。但该算法面临的挑战依然存在,主要难点有:①相似性计算耗时长。随着用户集的增大,逐个计算目标用户与其他用户之间相似度,将大大降低推荐算法的效率。②推荐算法虽然流行,但不够个性化。一般的推荐算法给用户推荐的商品大多数都是用户已经买过,或者大众普遍喜欢的。但是并不能很好地满足用户个性化需求。

为了解决协同过滤算法遇到的上述问题,本文提出了一种基于协调过滤与划分聚类的加权 TopN 推荐算法。该算法是对训练用户集划分聚类、改进用户相似性计算方法和推荐商品加权排行三方面进行改进。

2.1 训练用户集划分聚类

用户相似性在个性化推荐中至关重要,而随着用户集的增大,在计算相似度时耗时也增加。如何缩短计算时间成为研究的重点。为避免上述情况的发生,首先本文提出一个用户分类模型。在计算相似性之前,根据用户的属性信息(如性别、年龄、文化程度等)将用户集进行细化,分为不同的用户集合,如图 2 所示。

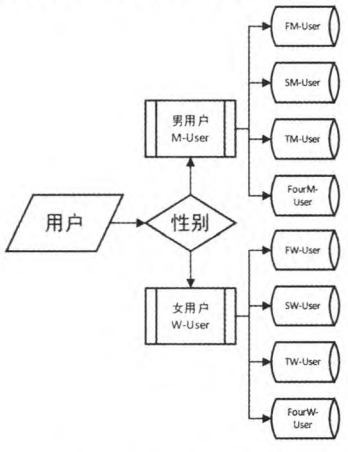

图 2 用户划分聚类

Fig. 2 User partition clustering

2.2 基于用户轨迹的 *POI* 推荐算法研究

利用协同过滤算法对目标用户产生推荐通常需要 3 个步骤[4-5]:①得到目标用户感兴趣的信息,如目标曾经去过的地点;②计算目标用户与用户集各用户间的相似度,并以此为依据得到与目标用户相似的用户集;③依据得到的用户集对目标用户进行推荐。由上可知相似性计算在整个算法中起到承上启下的作用,选择恰当的相似性计算方法能够明显提高整个推荐系统的推荐效率[6]。本文基于用户轨迹的 *POI* 推荐算法分为两大部分,一是相似性计算的改进;二是改进的加权 TopN 推荐算法。

2.2.1 相似性计算

传统的相似性算法,是将目标用户与用户集中用户逐个进行相似性计算。而实际情况中,目标用户 u 与用户集 v 中有很多用户并没有对同一 *POI* 产生过行为,即 $N(u) \cap N(v_j) = 0$,传统算法在该方面浪费了大量计算时间。针对该问题,本文对传统的相似性算法做了如下改进:

1)计算出 $N(u) \cap N(v_i) \neq 0$ 的用户对 (u, v_i);

2)对 $N(u)\cap N(v_i)\neq 0$ 的用户对,建立物品到用户的倒排表 $U1$,遍历目标用户得到其 POI,并保存对该 POI 产生行为的用户列表;

3)扫描倒排表中每个 POI 对应的用户列表,将用户列表中的两两用户对应的 $C[u][v_i]$ 加 1,假设目标用户 u 和用户 v_i 同时属于倒排表中 K 个 POI 对应的用户列表,则 $C[u][v_i]=K$。从而得到目标用户与其他用户之间不为 0 的 $C[u][v_i]$;

4)形成用户交集的矩阵 $C[u][v]=|N(u)\cap N(v)|$。同理,得到用户物品并集的矩阵 $C[u][v]=|N(u)\cup N(v)|$。

具体计算过程如图 3、图 4 所示。第一步,需要建立 POI - 用户的排序表;第二步,建立一个 $1\times n$ 的用户交集(并集)矩阵,对于物品 a,将 $W[A][B]$ 加 1,对于物品 b,将 $W[A][C]$ 加 1,以此类推;第三步,扫描完所有 POI 后,得到矩阵 $U1$($U2$),即相似度公式中的分子(分母)部分;第四步,将 $U1$ 除以分母 $U2$ 得到目标用户与该用户最终的兴趣相似度。

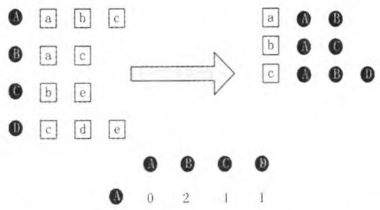

图 3　POI - 用户倒排表和分子矩阵 U1

Fig. 3　*POI* - user inverted list and molecular matrix U1

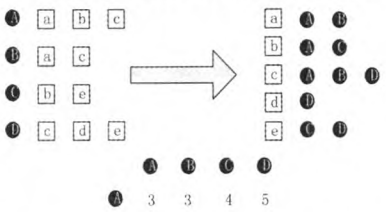

图 4　POI - 用户倒排表和分母矩阵 U2

Fig. 4　The user *POI* - inverted list and denominator matrix U2

2.2.2　加权 TopN 推荐算法

加权 TopN 推荐算法主要分为以下几个步骤:

首先,依据用户与目标用户之间的相似度,从大到小对用户进行排序,并存储于数组 SI 中,取数组 SI 中首个元素(最大相似度)并将其赋值给变量 Max$S$,依据公式(1)得到各个用户所涉及的 POI 所占权重,其中 K 为常数;

$$W[n]=\frac{SI[n]}{\mathrm{Max}S}\times K \tag{1}$$

然后,将不同用户的 POI 乘以相应权重,得到带有权重的 POI 集合,并将所有相关的 POI 进行叠加,并按照它们出现的次数进行排序,剔除目标用户涉及的 POI 点;

最后,利用 TopN 算法将前 n 个 POI 点推荐给目标用户。

该算法同一般的协同过滤推荐算法相比,基于用户轨迹的 POI 推荐算法更能满足目标用户的个性化需求。

3　实验与结果分析

3.1　实验数据

3.1.1　数据来源

选取不同年龄段且男女性别各 50 名志愿者,将手机 app 软件在后台实时记录用户的移动轨迹,通过用户两个月的真实活动来获得用户轨迹数据。POI 又称兴趣点,包含名称、类别、经纬度和附近的酒店、商铺等 4 方面信息。本文通过 Python 网络爬虫技术来获取郑州市区新浪微博 POI 签到数据,获取的时间段为从 POI 开始记录到 2015 年 12 月 13 日。

3.1.2　数据处理

本文实验数据处理阶段分为两部分:①压缩用户轨迹数据;②划分用户集。

用户轨迹数据的压缩。根据本文所提出的基于道路和 POI 的轨迹数据压缩算法,对用户轨迹进行压缩。并得到用户所涉及的 POI 单独存储于数据库中。

划分用户集。首先,根据属性信息对用户进行划分。本文依据用户的性别将用户划分为男、女两个集合,然后按照 18 岁以下、18 ~ 25 岁、26 ~ 35 岁、36 ~ 50 岁和 51 岁以上等 5 个年龄段细分为不同的用户集合。

3.2　评价标准

准确率:假设向目标用户推荐的 POI 个数为 T,其中有 TL 个是目标用户喜欢的。则推荐算法的准确率 $P=TL/T$。

召回率:是针对用户喜欢的总样本(Y)而言,它表示样本中有多少 POI 是被推荐给用户的。将样本中被推荐的 POI 记为 YL,则推荐算法的召回率 $R=YL/Y$。

流行度:是指评测推荐结果的新颖度,本文用推荐列表中物品的平均流行度度量推荐结果的新颖度。如果推荐的物品都很热门,说明推荐的新颖度较低;否则,说明推荐结果比较新颖。

3.3　实验方案

依据本文提出的基于用户轨迹的 POI 推荐算法,该实验主要分为以下几个步骤。

1)搜索与目标用户属性相同的用户集合 A。根据目标用户的属性信息,从数据库中找到与目标用户属性相同的用户集合,并得到其相关的 POI;若目标用户未注册相关用户信息则默认为全部用户数据,如图 5 所示。

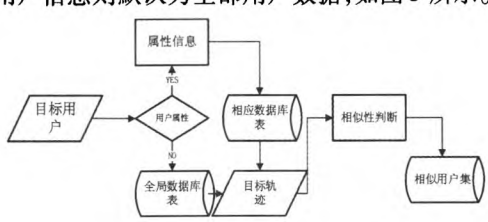

图 5　推荐算法实验模型

Fig. 5　Experimental model of recommendation algorithm

2)从集合 A 中寻找与目标用户交集不为空的用户集合 B,获取相应的 POI。并按照本文相似性计算算法,得

到目标用户与集合 B 中各用户的相似度。

3)依据用户与目标用户之间的相似度,从大到小对用户进行排序,并存储于数组 SI 中。取数组 SI 中首个元素(最大相似度)并将其赋值给变量 Max$S$,依据公式(1)得到各个用户所涉及的 POI 所占权重,并将取值为 100。最后,将不同用户的 POI 乘以相应权重,得到带有权重的 POI 集合。再将所有相关的 POI 进行叠加,并按照它们出现的次数进行排序,剔除目标用户涉及的 POI 点。最后利用 TopN 算法得到将前 10 个 POI 点推荐给目标用户。

3.4 结果分析

为了检验本文基于用户轨迹的 POI 推荐算法(简称 UserTN)及基于道路和 POI 的轨迹数据压缩算法的效率。本文共做了两个实验:实验 1,利用传统的协调过滤推荐算法和划分后的用户集合;实验 2,利用本文改进的推荐算法 UserTN 和划分后的用户集合。

表 1 中给出了实验 1 和实验 2 在离线状态下的相关性能。从表 1 中可知,UserTN 算法在准确率、召回率等方面都优于传统的协同过滤推荐算法,虽然在流行度方面略逊于 UserCF,但是 POI 的流行度并不能反映用户的满意度,恰恰是跟目标用户最相似的用户 POI 满足其个性化需求。

表 1 实验 1 和实验 2 的性能对比

Tab. 1 Performance comparison of experiment 1 and experiment 2

	准确率(%)	召回率(%)	流行度(%)
实验一	25.20	12.17	7.289 817
实验二	30.42	14.20	4.315 728

4 结束语

本文研究了基于用户轨迹的 POI 个性化推荐算法和模型。针对研究目的,本文提出了一种基于道路和 POI 的轨迹数据压缩算法,高效地压缩了用户轨迹数据,并快速、准确地提取出用户涉及的 POI。而本文提出的基于用户轨迹的 POI 推荐算法,经实验证明:与传统的协同过滤推荐算法相比,提出的算法较大程度上提高了推荐结果的个性化,在准确率和召回率方面也优于传统的协同过滤推荐算法。

参考文献:

[1] Muckell J, Hwang J H, Lawson C T, etal. Algorithms for compressing GPS trajectory data: an empirical evaluation [C]//Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geogrephic Information Systems, ACM, New York; ACM, 2010:402 - 405.

[2] 吴颜,沈洁,顾天竺,等. 协同过滤推荐系统中数据稀疏问题的解决[J]. 计算机应用研究,2007,24(6):94 - 97.

[3] 赵博,赵鹏飞. 推荐算法综述[J]. 山西大学学报:自然科学版,2011,34(3):337 - 350.

[4] Breese J S, Hecherman D, Kandie C. Empirical anilysis of predictive algorithms for collaborative filtering[C]// UAI98:Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence San Francisco CA:Morgan Kaufnann, 1998:43 - 52

[5] Su xiaoyuan, Khoshgoftaar TM. A survey of collaborative filtering techniques[J/OL][J]. Advances in Artificial Intelligence,2009,1(1):421 - 525.

[6] 嵇晓声,刘宴兵,罗来明. 协同过滤中基于用户兴趣度的相似性度量方法[J]. 计算机应用,2010,30(10):2 618 - 2 620.

[7] 任看看,钱雪忠. 协同过滤算法中的用户相似性度量方法研究[J]. 计算机工程,2015,41(8):18 - 22.

[8] 陈康,黄晓宇,王爱宝,等. 基于位置信息的用户行为轨迹分析与应用[J]. 电信科学,2013,12(4):118 - 124.

[9] 陈克寒,韩盼盼,吴健. 基于用户聚类的异构社交网络推荐算法[J]. 计算机学报,2013,36(2):349 - 359.

[10] 项亮. 推荐系统实践[M]. 北京:人民邮电出版社,2012.

[11] 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,2010,6(6):23 - 30.

[12] 张达夫,张昕明. 基于时空特征的 GPS 轨迹数据压缩算法[J]. 交通信息与安全,2013,31(3):6 - 9.

[13] 李伟,陈毓芬,李萌,等. 基于情境的 POI 个性化推荐方法研究[J]. 武汉大学学报:信息科学版,2015,40(6):829 - 833.

[编辑:栾丽杰]

(上接第 54 页)

[15] Cheng Y C, Chen S Y. Image classification using color, texture and regions[J]. Image and Vision Computing, 2003(21):759 - 776.

[16] 杨耘,徐丽,颜佩丽. 条件随机场框架下基于随机森林的城市土地利用/覆盖遥感分类[J]. 国土资源遥感,2014,26(4):51 - 55.

[17] Tang Y Q, Zhang L P, Huang X. Object - oriented change detection based on the kolmogorov - smirnov test using high - resolution multispectral imagery[J]. International Journal of Remote Sensing,2011,32(20):5 719 - 5 740

[18] 邓劲松,李君,王珂. 基于多时相 PCA 光谱增强和多源光谱分类器的 SPOT 影像土地利用变化检测[J]. 光谱学与光谱分析,2009,29(6):1 627 - 1 631.

[19] Wang A P, Wang S G, LucieerA. Segmentation ofmultispectral high - resolution satellite imagery based on integrated feature distributions[J]. International Journal of Remote Sensing,2010,31(6):1 471 - 1 483.

[20] Ojala T, Pietikäinen M. Unsupervisedtexture segmentation using feature distributions[J]. Pattern Recogniton,1999, 32(3):477 - 486.

[编辑:张 曦]