基于模糊聚类的旅游推荐算法

张应辉,李 雪
(东北大学 计算机科学与工程学院,辽宁 沈阳 110000)

摘 要:在旅游领域中,旅游者常常在旅游前从互联网上获取所需信息,但是在线旅游业日益严重的信息过载现象,使得用户不能得到他们想要的个性化信息。传统的基于协同过滤的旅游推荐研究普遍都存在稀疏性和可扩展性等问题,基于知识的推荐研究有时因用户无法表达清楚他们的需求而无法得到满意的推荐。针对已有的旅游推荐算法存在的问题,提出了一种基于模糊聚类的旅游推荐算法,为用户推荐符合其需求和偏好的旅游产品。该算法利用标签构建用户偏好景点模型和景点特征属性模型,对数据集进行模糊聚类,同时提出新的相似度度量。在此基础上,组合基于内容和协同过滤技术进行混合推荐。实验结果表明,该算法能显著提高推荐系统的效率以及可扩展性和准确度。

关键词:个性化;标签;相似性度量;模糊聚类;混合推荐
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2016)12-0099-04
doi:10.3969/j.issn.1673-629X.2016.12.022

A Tourism Recommendation Algorithm Based on Fuzzy Clustering

ZHANG Ying-hui, LI Xue
(School of Computer Science and Engineering, Northeastern University, Shenyang 110000, China)

Abstract: In the field of tourism, tourists often get the information they need on the Internet before traveling, but the phenomenon of information overload online in tourism industry is becoming more and more serious, so that personalized information cannot be obtained by users. The problems of sparsity and scalability exist in the traditional tourism recommendation algorithm based on collaborative filtering, and sometimes users can't express their needs and can't be satisfied with the recommendation based on the knowledge of the recommendations. For these problems, a tourism recommendation algorithm based on fuzzy clustering is proposed, which is used for the users to recommend the tourism products that meet their needs and preferences. Tags are used by the algorithm to build user's preference models and sights feature attribute model, fuzzy clustering on them. A new similarity measure is proposed. On this basis, the combination of content-based and collaborative filtering technology is recommended. Experimental results show that the proposed algorithm can significantly improve the efficiency, scalability and accuracy of the recommendation system.

Key words: individualization; tags; similarity measurement; fuzzy clustering; hybrid recommendation

0 引 言

旅游推荐算法[1-3]的研究是旅游领域研究的热点。旅游网站不断兴起,推荐精度的高低直接影响用户是否选择预定该线路,影响用户对该旅游网站信息的兴趣度,兴趣度的高低决定了用户对该旅游网站的使用率。针对个性化旅游推荐[4-6]问题,学者们进行了深入研究。例如,基于协同过滤技术的旅游推荐研究[7]在一定程度上提高了推荐的多样性,但是普遍都存在稀疏性和可扩展性的问题。基于知识的、会话式的旅游推荐[8]方式使用交互 & 个性化代理以会话的形式逐步发现用户的偏好和需要,然后利用多属性效用理论对推荐结果进行排序,一定程度上提高了推荐的精确度。但此方法需要大量的领域知识和推理技术,需要考虑多方面的因素,有时用户很难准确地表达自己的需求,推荐效率缓慢。

针对上述问题,提出一种基于模糊聚类[9-12]的旅游推荐算法(Tourism Recommendation algorithm Based on Fuzzy Clustering, TRBFC),建立了新的用户偏好景

收稿日期:2016-06-03 **修回日期**:2016-09-08 **网络出版时间**:2016-11-22
基金项目:国家自然科学基金资助项目(61262058)
作者简介:张应辉(1972-),男,副教授,硕士生导师,研究方向为计算机图像处理、机器学习;李 雪(1991-),女,硕士研究生,研究方向为数据挖掘、机器学习。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20161122.1227.020.html

点模型[13-14],提出了一种新的相似度计算方法,使用模糊聚类的方法对数据集进行聚类,在此基础上,组合基于内容和协同过滤的技术进行混合推荐。该算法使得系统的推荐效率、可扩展性进一步提高,改善了系统的稀疏性,进一步提高了推荐的准确率。

1 相关定义

TRBFC 算法在构建用户偏好景点模型时,主要考虑了用户使用过的景点标签。当用户浏览旅游网站时,用户喜欢的景点都会有相应的标签,比如 Tom 喜欢的景点标签中经常出现“主题”、“海边”等短语,那么他可能喜欢主题游或海边游,其中“主题”出现的频率较高,Tom 可能更喜欢此类景点。

定义 1:如果系统中有 q 类景点标签,那么对用户,通过 TRBFC 算法构建的用户景点偏好模型如式(1)所示:

$$O_{u_i} = (p_1, p_2, \cdots, p_q) \tag{1}$$

其中,p_q 表示标签 q 被用户 u_i 使用的频率(即次数)。

定义 2:如果系统中有 q 类景点标签,那么对于景点,通过 TRBFC 算法构建的景点特征属性模型如式(2)所示:

$$I_{s_i} = (a_1, a_2, \cdots, a_q) \tag{2}$$

其中,a_q 表示标签 q 是否是景点 s_i 的标签。

定义 3:用户-标签矩阵 $\boldsymbol{B}$(user-tags frequency matrix)。如果有 m 个用户 $U_1 = (u_1, u_2, \cdots, u_m)$ 和 q 个标签 $T_1 = (t_1, t_2, \cdots, t_q)$,它们之间形成一个 $m * q$ 的矩阵,其中行为用户,列为标签,如下所示:

$$\boldsymbol{B} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mp} \end{bmatrix} \tag{3}$$

其中,x_{ij} 表示用户 u_i 使用标签 t_i 的个数。

定义 4:景点-标签属性矩阵 $\boldsymbol{S}$。如果有 n 个景点 $I = (s_1, s_2, \cdots, s_n)$ 和 q 个标签 $T = (t_1, t_2, \cdots, t_q)$,它们之间形成一个 $n * q$ 的矩阵,其中行为景点,列为标签,如下所示:

$$\boldsymbol{S} = \begin{bmatrix} y_{11} & \cdots & y_{1q} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nq} \end{bmatrix} \tag{4}$$

其中,y_{ij} 表示景点 s_i 是否包含标签 t_i,包含则值为 1,反之为 0。

定义 5:用户评分矩阵 $\boldsymbol{R}$(user-item matrix)。如果有 m 个 $U_1 = (u_1, u_2, \cdots, u_m)$ 和 n 个景点 $I = (s_1, s_2, \cdots, s_n)$,它们之间形成一个 $m * n$ 的矩阵,其中行为用户,列为景点,如下所示:

$$\boldsymbol{R} = \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{bmatrix} \tag{5}$$

其中,r_{ij} 表示用户 u_i 对景点 s_i 的评分。

评分值为[1,5]之间的整数,评分值由高到低表明用户对该景点兴趣的高低。若未评分,则取值 0。

定义 6:欧氏距离。欧几里德距离又叫欧氏距离,常用来计算两个向量间的距离,并认为这是两个向量的差距。TRBFC 算法采用欧氏距离,如式(6)所示:

$$d_{ti} = \sqrt{\sum_{j=1}^{q} (x_{tj} - x_{ij})^2} \tag{6}$$

其中,d_{ti} 表示用户 u_t 对用户 u_i 偏好景点之间的距离;x_{ij} 为定义 3 中矩阵 $\boldsymbol{B}$ 中用户使用标签的频率(即个数)。

2 TRBFC 算法的实现

由于一个景点可能拥有多个标签,可以属于多个不同的类,所以首先采用模糊聚类的方法对用户-标签数据集和景点-标签数据集进行聚类,使相近的景点或用户分为一组,其次组合基于内容和协同过滤的推荐算法,按照一定的关系组合二者,进行旅游景点的推荐。

首先对 O_{u_i} 进行模糊聚类。

(1)基本参数初始化。聚类的最终类别个数 c,$2 \leqslant c < n$,n 是标签个数;加权指数 m,取[1.5,2.5]间的值更好[15];迭代停止阈值 θ($\theta > 0$);用户-标签频率矩阵 $\boldsymbol{B}$;聚类中心矩阵 $\boldsymbol{V}^0$ 和迭代次数计数器 $f = 0$。

(2)用户-标签隶属度矩阵 $\boldsymbol{U}^f$ 的更新。用式(7)进行更新:

$$u_{ij}^f = \begin{cases} \left\{ \sum_{k=1}^{c} \frac{d_{ij}^f}{d_{kj}^i} \right\}^{-1}, d_{ij} > 0 \\ 1, d_{ij} = 0 \\ 0, d_{ij} = 0 \text{ 且 } j \neq 0 \end{cases} \tag{7}$$

其中,d_{ij}^f 表示 f 次迭代后用户 u_i 与第 j 个聚类中心之间的距离,由式(6)计算。隶属度越大,相似度越高。

(3)用户-标签聚类中心矩阵 $\boldsymbol{V}^{f+1}$ 更新,使用式(8):

$$v_i^{f+1} = \frac{\sum_{k=1}^{n} (u_{ik}^{f+1})^m \times o_{u_k}}{\sum_{k=1}^{n} (u_{ik}^{f+1})^m}, i = 1, 2, \cdots, c \tag{8}$$

其中,u_{ij}^f 表示迭代 f 次后属性度 $\boldsymbol{U}$ 中的元素。

(4)如果 $\| \boldsymbol{V}^f - \boldsymbol{V}^{f+1} \| < \theta$,则算法停止并返回用户-标签隶属度矩阵 $\boldsymbol{U}$ 和用户-标签聚类中心矩阵 $\boldsymbol{V}$,否则 $f = f + 1$,转向步骤(2)进行迭代计算。

(5)对于目标用户 u_t,根据隶属度找到它所在的类别,把式(6)作为新的相似度度量,计算 u_t 与其所在类别中其他用户之间的相似度,按照相似度大小排序,排在最前面的 N 位即可作为目标用户的邻居集,记为 $N(u_t)$。

同理可以对 I_{s_i} 模糊聚类后获取景点-标签隶属度矩阵 $\boldsymbol{I}$ 和景点 - 标签聚类中心矩阵 $\boldsymbol{Q}$。此处不再证明。

在此基础上,推荐结果由基于内容和协同过滤的混合推荐算法来推荐获得。

(1)使用协同过滤方法对于目标用户 u_t 的邻居集 $N(u_t)$,结合式(5)给出的用户评分矩阵 $\boldsymbol{R}$,对目标用户未选择的景点做预测评分,如式(9) 所示:

$$r_{tw} = \overline{r_t} + \frac{\sum_{u_i \in N<u_i>} \left|1 - \frac{1}{d_{ti}}\right| (r_{iw} - \overline{r_i})}{\sum_{u_i \in N<u_i>} \left|1 - \frac{1}{d_{ti}}\right|} \tag{9}$$

其中,r_{tw} 表示目标用户 u_t 对景点 w 做的预测评分;d_{ij} 的值应该大于等于 1。

得到预测评分后,按其高低把获得预测评分最高的 Top-K 个项目放入一个集合 M 中。

(2)使用基于内容的方法,根据隶属度判断目标用户 u_t 正在查看的或者已经存在景点 s_i 所在的模糊类类别。s_i 可能属于多个类。利用式(10)计算目标用户 u_t 与所属聚类类别中其他景点的相似性:

$$\mathrm{sim}(a, s_i) = \frac{2 \times |\mathrm{key}(a) \cap \mathrm{key}(s_i)|}{|\mathrm{key}(a)| + |\mathrm{key}(s_i)|} \tag{10}$$

其中,$\mathrm{sim}(a, s_i)$ 是由景点 s_i 和类中其他景点 a 之间标签相同的个数比两者标签总的个数所得。

设置一个集合 H,一个阈值 β,当 $\mathrm{sim}(a, s_i) > \beta$,把景点 s_i 放入 H 中。对集合中的景点按相似度值大小排序。获取 top-N 个景点的推荐集合,N 的值取 5。

(3)综合集合 M 和 H 中的景点,两个集合相交得到最终的景点推荐集合 HM。

3　实验结果及分析

3.1　数据来源

使用从途牛网中获取的旅游景点信息进行实验。在选取的整个数据集中,所有的景点数据为 512,景点评分数据为 67 690,评分取[1,5]中的任意整数,评分值由高到低代表旅游者对该景点的满意程度。在得到的数据集中,以用户-标签矩阵为例,形式如表 1 所示。

3.2　评价指标

训练集由随机抽取 50 000 条景点评分组成,测试集由剩余的数据组成,分别用传统的基于知识的旅游推荐、基于用户的协同过滤方法和改进算法进行比较。采用准确率和召回率作为评测标准。

表 1　用户-标签矩阵 $\boldsymbol{B}$

标签用户	跟团游	自助游	国内游	……	海边游
用户 1	5	2	1	……	4
用户 2	3	3	5	……	1
……	……	……	……	……	……
用户 m	2	4	3	……	3

$$P_u = \frac{\sum_{u \in U_1} |\mathrm{TM}_u \cap T_u|}{\sum_{u \in U_1} |\mathrm{TM}_u|} \tag{11}$$

$$R_u = \frac{\sum_{u \in U_1} |\mathrm{TM}_u \cap T_u|}{\sum_{u \in U_1} |T_u|} \tag{12}$$

其中,P_u 为准确率;R_u 为召回率;TM_u 为算法用户推荐景点的集合;T_u 为用户在测试集上喜欢的景点的集合。

3.3　实验分析

图 1 和图 2 分别为传统的基于知识推荐、协同过滤推荐和文中算法的准确率测试和召回率测试。其中,U_1 是最终推荐景点数目为 10 的数据集,$U_2 \sim U_5$ 分别是最终推荐景点数目为 15、20、25、30 的数据集,当最终推荐景点数目达到 30 时,准确率值上升缓慢,所以最终推荐景点数目不宜选择过大。

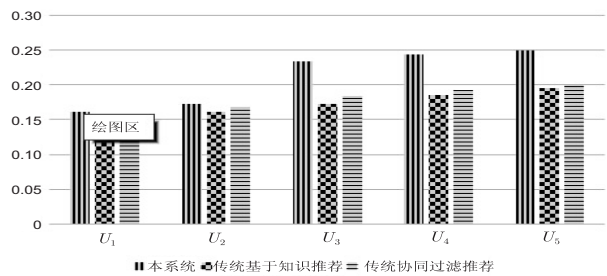

图 1　准确率测试

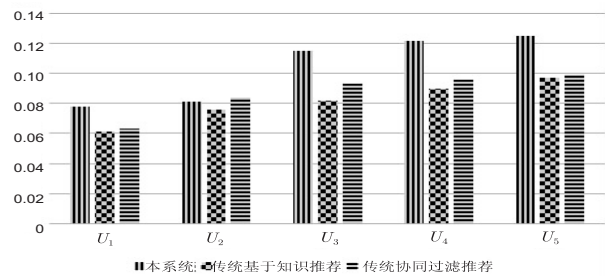

图 2　召回率测试

从两图中可以看出,与传统的算法相比,数据模糊聚类后,混合推荐算法的推荐精度要高一些。

4　结束语

针对传统旅游推荐算法推荐效率不高的问题,提

出了一种模糊聚类方法,采用新的相似度度量,在一定程度上缩短了寻找用户邻居集和相似景点的时间,提高了效率和扩展性。另外采用混合推荐技术,在一定程度上改善了推荐系统的稀疏性和冷启动问题。

参考文献:

[1] Hwang S, Yan W. On-tour attraction recommendation in a mobile environment[C]//IEEE conference on pervasive computing and communications. New Jersey: IEEE Press, 2012.

[2] Ricci F, Rokach L, Shapira B, et al. Recommender system handbook[M]. [s. l.]: Springer, 2011.

[3] 陈 梅. 旅游信息智能推荐系统的研究与设计[D]. 贵阳: 贵州大学, 2010.

[4] Liu Q, Ge Y, Li Z M, et al. Personalized travel package recommendation[C]//IEEE international conference on data mining. New Jersey: IEEE Press, 2011: 407-416.

[5] 安 维, 刘启华, 张李义. 个性化推荐系统的多样性研究进展[J]. 图书情报工作, 2013, 57(20): 127-135.

[6] 胡纳纳, 李琳琳, 武 尚. 个性化的旅游推荐系统[J]. 信息技术, 2013(2): 135-139.

[7] 侯新华, 文益民. 基于协同过滤的旅游景点推荐[J]. 计算机技术与自动化, 2012, 31(4): 116-119.

[8] 王显飞, 陈 梅, 李小天. 基于约束的旅游推荐系统的研究与设计[J]. 计算机技术与发展, 2012, 22(2): 141-145.

[9] Zenebe A, Zhou Lina, Norcio A F. User preferences discovery using fuzzy models[J]. Fuzzy Sets and Systems, 2010, 161: 3044-3063.

[10] Srivastava V, Tripathi B K, Pathak V K. An evolutionary fuzzy clustering with Murkowski distances[C]//Proceedings of the 2011 international conference on neural information processing. Shanghai, China: [s. n.], 2011.

[11] Zhang Chen, Liu Bing. Possibilistic fuzzy clustering algorithm based on sample weighted[C]//Proceedings of 3rd international workshop on intelligent systems and applications. Wuhan, China: [s. n.], 2011.

[12] Tsai Du-Ming, Lin Chung-Chan. Fuzzy C-means based clustering for linearly and nonlinearly separable data[J]. Pattern Recognition, 2011, 44(8): 1750-1760.

[13] Huang Weidong, Khoury R, Dawborn T, et al. WeBeVis: analyzing user web behavior through visual metaphors[J]. Science China Information Sciences, 2013, 56(5): 1-15.

[14] Wu Xiyuan, Zheng Qinghua, Wang Ping. A intelligent method of modelling web user interest[J]. Journal of New Industrialization, 2014(9): 39-43.

[15] 肖曼生, 阳姊兰, 张居武, 等. 基于模糊相关度的模糊 C 均值聚类加权指数研究[J]. 计算机应用, 2010, 30(12): 3388-3390.

(上接第 98 页)

模技术[J]. 测绘科学, 2011, 36(1): 213-214.

[2] 王星捷, 李春花. 基于 Unity3D 平台的三维虚拟城市研究与应用[J]. 计算机技术与发展, 2013, 23(4): 241-244.

[3] 曹兆峰, 何燕兰, 李胜才. 基于 Sketchup 和 ArcGIS 的数字城市三维建模技术[J]. 地理空间信息, 2014, 12(5): 46-47.

[4] 孙 钊, 吴志华, 熊 伟. 基于三维数字技术的城市设计研究与应用[J]. 城市规划学刊, 2009(7): 239-241.

[5] 赵子龙. 基于 3ds Max 的城市三维建模技术[J]. 价值工程, 2013, 32(4): 184-185.

[6] 万宝林. 3DS MAX 与 SketchUp 的三维城市建模技术实验对比分析[J]. 测绘地理信息, 2015, 40(2): 23-25.

[7] 李 娟, 吴红梅, 陈永波. 基于 Skyline 的三维数字城市建设项目的设计与研究[J]. 测绘与空间地理信息, 2015, 38(10): 165-167.

[8] 宋宜容, 严康文. 基于 GoogleEarth 的三维数字浏览系统的设计与实现[J]. 湖北大学学报: 自然科学版, 2015, 37(2): 107-111.

[9] Tong L, Li Yanlin. Research progress of three-dimensional digital model for repair and reconstruction of knee joint[J]. Chinese Journal of Reparative & Reconstructive Surgery, 2013, 27(1): 50-53.

[10] Zhang Qiuwen, Wang Cheng, Shi Zhongchao, et al. A three dimensional modeling and simulation platform design for digital city[J]. Proc Spie, 2005, 6(3): 59855S.

[11] Bremer M, Mayr A, Wichmann V, et al. A new multi scale 3D-GIS-approach for the assessment and dissemination of solar income of digital city models[J]. Computers Environment and Urban Systems, 2016, 57: 144-154.

[12] Sharma S A, Agrawal R, Jayaprasad P, et al. Development of '3D city models' using IRS satellite data[J]. Journal of the Indian Society of Remote Sensing, 2015, 23: 1-10.

[13] Xiong B, Jancosek M, Elberink S O, et al. Flexible building primitives for 3D building modeling[J]. ISPRS Journal of Photogrammetry & Remote Sensing, 2015, 101: 275-290.

[14] McDermid R M, Alatalo K, Blitz L, et al. The Atlas3D Project-XXX. Star formation histories and stellar population scaling relations of early-type galaxies[J]. Monthly Notices of the Royal Astronomical Society, 2015, 448(4): 3484-3513.