

基于领域特征值的协同过滤个性化推荐方法

方超, 暴建民, 薛四猛

(南京邮电大学 物联网学院, 江苏 南京 210003)

摘要: 知识发现领域中, 个性化推荐技术因其应用广泛受到了业界的广泛关注和高度重视。但由于用户隐私保护方面的限制, 现有的推荐系统不能直接挖掘用户的个人信息, 因此只能采用表征用户爱好的特征值来间接地挖掘用户信息。针对此类问题, 提出了一种新的推荐方法。该方法可自动提取相应领域的特征值, 并基于领域关键词过滤冗余的领域特征值, 从而据此构建用户偏好模型, 并与协同过滤算法绑定, 生成最终的推荐结果。为验证所提出推荐方法的有效性和可行性, 基于实时数据集与其他已有的推荐方法进行了对比实验, 并基于对比实验结果进行了相关的分析研究。对比验证实验结果及其分析表明, 该推荐方法能够有效地提取领域特征值, 其推荐的精准度也有所提高。

关键词: 领域特征值; 协同过滤; 用户偏好模型; 个性化推荐

中图分类号: TP301

文献标识码: A

文章编号: 1673-629X(2017)11-0088-04

doi: 10.3969/j.issn.1673-629X.2017.11.019

A Personalized Collaborative Filtering Recommendation Method Based on Domain Features

FANG Chao, BAO Jian-min, XUE Si-meng

(College of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: In knowledge discovery, personalized recommendation technology has received extensive concern and high attention because of its wide application. However, due to the limitations of user privacy protection, the existing recommendation system can't directly mine the user's personal information. So, the features which imply user preference to indirectly mine user information can be utilized. In order to solve above problem, a new recommendation method is proposed which can automatically extract relevant domain features and filter the redundant domain features based on domain keywords to construct a user preference model and generate the final recommendation result in combination with the collaborative filtering algorithm. To verify its effectiveness and feasibility, compared with other existing recommendation methods based on a real time data sets the experiments for verification are conducted. The results of contrast experiments and relevant analysis show that it can effectively extract the domain features and its accuracy of the recommendation is improved.

Key words: domain features; collaborative filtering; user preference model; personalized recommendation

0 引言

当今社会正处于一个数据爆发式增长的时代。由于信息的增长, 用户得到他们想要的有用信息越发困难^[1], 这种信息过载的问题越来越严重。因此个性化推荐系统应运而生, 它利用数据挖掘和机器学习技术从大数据中挖掘用户的需求与偏好, 并为用户提供精确的物品推荐^[2]。

近年来, 在学术与工业领域开发了许多推荐系统。其中协同过滤算法被广泛应用于推荐系统中, 是最好的推荐算法之一^[3-8]。协同过滤算法主要使用邻近技术计算用户与用户之间的联系, 然后预测目标用户对

物品的偏好, 物品的预估值是用这个偏好的最近邻居权重值来表示。最后, 推荐系统为目标用户推荐物品是利用这个预估值来推荐。协同过滤的好处是它对推荐目标没有特殊要求, 并且能有效地处理复杂的非结构化目标, 例如书籍或者电影。但是, 其性能也存在数据的稀疏性、冷启动以及可扩展性等诸多限制。

为了解决协同过滤算法存在的问题, 学者们提出了许多改进方法来增强推荐系统的性能。Hu L 等提出了一种基于物品特征值与用户偏好的混合协同过滤推荐算法^[9], 该算法提高了推荐系统的精确性, 并能更加容易地处理数据的稀疏性。Jung S. Y 等提出了一

收稿日期: 2016-11-19

修回日期: 2017-03-10

网络出版时间: 2017-07-19

基金项目: 国家自然科学基金资助项目(61100213); 南京邮电大学教育部重点实验室开放研究基金(ZS035NY11005)

作者简介: 方超(1991-), 男, 硕士研究生, 研究方向为数据挖掘; 暴建民, 正高级工程师, 硕士生导师, 研究方向为物联网、大数据。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20170719.1112.076.html>

种用户偏好模型,有效缓解了数据的稀疏性,提高了预测结果的精确性^[10]。Choi S H 等提出了混合推荐方法以减少大数据量^[11],该方法通过计算最远用户间的距离以减少数据集的规模,成功避免了大数据带来的可扩展性与稀疏性的问题。

此外,用户偏好是个性化应用的关键因素,是个性化推荐系统的本质所在。然而,查询用户显性的信息很困难,不能直接挖掘出用户偏好。相反,能查询用户隐性的信息^[12]。所以,研究者们致力于开发隐性启发方法挖掘用户记录与行为信息,从中得到有效的用户偏好集。Chen Y Y 等针对个性化的旅游推荐^[13],从照片内容中挖掘人们的基本属性与旅游类型信息,以此构建用户偏好集。该方法提高了推荐的精度。Zhang J 等提出了一种新的推荐方法,从物品及其评价信息中挖掘物品的特征,构建用户偏好模型^[14],但没有很好地解决用户隐私问题。

同时,物品的特征在不同的领域内是不同的,因此开发了有针对性的领域推荐系统。Anand S T 等提出了书籍推荐系统,将内容过滤的特征、协同过滤算法以及相关规则挖掘绑定^[15]。Chen J H 等提出了混合过滤算法,提供了多功能的旅游信息^[16]。Chen R C 等提出一个糖尿病医药推荐系统,采用了基于医院专家提供的知识领域本体^[17]。

在此,文中提出一种新的推荐方法,其能够自动提取物品的领域特征值,并通过领域关键词验证过滤冗余的领域特征值,根据得到的领域特征值集构建用户偏好模型,将用户偏好模型与协同过滤算法绑定产生推荐结果。

1 相关理论

1.1 领域特征值自动提取

1.1.1 数据预处理

一般来说,大多数已存在的数据库文件以网页页面的形式存在,例如 HTML 或者 XML 标签。因此,必须过滤这些标签,得到需要的数据集。使用一个实时的数据集,在数据预处理前,采用开放网页爬虫软件 Hertrix 收集用户数据,并存储在数据库中。然后对这些数据进行格式化,得到下一步特征值提取时输入的标准数据集。

1.1.2 领域特征值提取

图 1 是特征值提取的整个框架流程。

首先,使用解析器(FudanNL-Process)解析所有的句子,得到词袋。然后,通过验证简单的名词与动词组来帮助找到能明确表达领域特征值的词。最后,利用关联规则挖掘领域中所有相关的特征值,得到领域特征值集。然而,领域特征值集可能有许多用户并不感

兴趣或者存在冗余的特征值,所以,需要通过清洗特征值移除不准确的特征值。领域关键词是在某一个领域中描述物品特征值的词。它能够进一步清洗特征值,过滤冗余特征值。在验证领域特征值之后,将得到最终的物品领域特征值集。

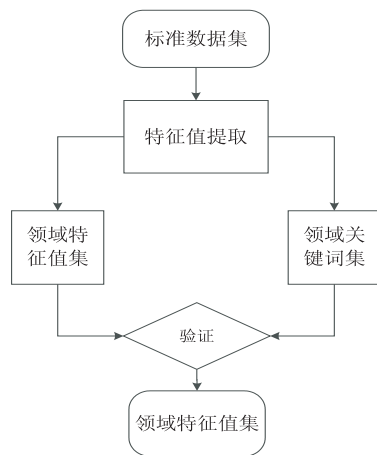


图 1 领域特征值提取流程

将用户的领域特征值集合定义为 $D = \{D_1, D_2, \dots, D_M\}$, 其中 $D_M (1 \leq M < l)$ 是第 M 个领域特征值, l 是被选择的特征值数量。

1.2 推荐引擎

1.2.1 用户偏好模型构建

在个性化推荐系统中,物品与用户是其两个主要的实体。将物品集合定义为 $I = \{I_1, I_2, \dots, I_l\}$, 用户集合定义为 $U = \{U_1, U_2, \dots, U_j\}$ 。

在上一节,从用户信息中提取的领域特征值集定义为 $D = \{D_1, D_2, \dots, D_M\}$ 。例如,酒店领域特征值包括环境、服务、价格、区域等等,用户偏好可以被物品的特征值所反映。所以,定义 $F_j = \bigcup_{i \in I_j} D_i = \{f_{j1}, f_{j2}, \dots, f_{jr}\}$ 去描述用户偏好领域特征值集。其中, F_j 表示用户 j 选择的物品特征值集合; f_{jr} 表示用户 j 喜爱的物品的特征值集合。

同时,定义物品评分集为 $S_j = \{s_{ji} \mid j \in U, i \in I_j, I_j \subseteq I\}$ 。由物品评分集 S_j 与特征值集 F_j , 得到用户 j 的偏好矩阵,如下:

$$X_m = (e_{NM}^j) = (S_j \cdot F_j) = \begin{pmatrix} s_{11}f_{11} & s_{12}f_{12} & \cdots & s_{1r}f_{1r} \\ s_{21}f_{21} & s_{22}f_{22} & \cdots & s_{2r}f_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ s_{j1}f_{j1} & s_{j2}f_{j2} & \cdots & s_{jr}f_{jr} \end{pmatrix} \quad (1)$$

然后,计算用户 j 的偏好向量。

$$\vec{W}_j = [w_{j1}, w_{j2}, \dots, w_{jm}, \dots, w_{jr}] \quad (2)$$

其中, w_{jm} 表示 f_{jm} 的权重,满足 $\sum_{m=1}^r w_{jm} = 1, 0 \leq w_{jm} \leq 1$ 。已推荐用户与历史用户的偏好权重向量分别

定义为 \vec{W}_R 和 \vec{W}_P 。

1.2.2 相似度计算

使用余弦相似度计算已推荐用户与历史用户的相似度。

$$\text{sim}(R, P) = \cos(\vec{W}_R, \vec{W}_P) = \frac{\vec{W}_R \cdot \vec{W}_P}{\|\vec{W}_R\|_2 \times \|\vec{W}_P\|_2} = \frac{\sum_{h=1}^n \vec{W}_{R,h} \times \vec{W}_{P,h}}{\sqrt{\sum_{h=1}^n \vec{W}_{R,h}^2} \sqrt{\sum_{h=1}^n \vec{W}_{P,h}^2}} \quad (3)$$

其中, \vec{W}_R 和 \vec{W}_P 分别为已推荐用户和历史用户的偏好权重向量; $\vec{W}_{R,h}$ 为第 h 个已推荐用户偏好的权重; $\vec{W}_{P,h}$ 为第 h 个历史用户偏好的权重。

1.2.3 产生推荐

为已推荐用户,采用一个权重平均值的方法来预估物品预测分数 p_{ji} 。

$$p_{ji} = \bar{r} + K \sum_{R_j \in A} \text{sim}(R, P) \times (r_j - \bar{r}) \quad (4)$$

$$K = 1 / \sum_{R_j \in A} \text{sim}(R, P) \quad (5)$$

其中, $\text{sim}(R, P)$ 为已推荐用户与历史用户的相似度; A 为过滤之后的特征值集; K 为一个标准化因子; r_j 为历史用户偏好的评分; \bar{r} 为定义为候选产品的平均评分。

得到的物品预测评分按照从大到小依次排序,将序列的前 N 项物品推荐给目标用户。

2 实验

2.1 实验数据集和评价标准

为了预估推荐方法的精确度,使用一个实时的数据集。采用的领域数据集是在著名的旅游网站(www.tripadvisor.com)上抓取的,主要收集了网站上北京酒店的数据。得到了 2 467 个用户与 245 个酒店的 213 566 条记录,其中 80% 的数据用来训练模型,20% 的数据用来测试方法性能。

利用归一化平均绝对差(NMAE)、精确度(Precision)这两个指标来预估提出的推荐方法的性能。

NMAE 对预测的准确性进行预估,计算公式为:

$$\text{NMAE} = \frac{1}{\Delta r M} \sum_{j=1}^M |r_j - \text{pr}_j| \quad (6)$$

其中, r_j 为用户 j 的预测评分; pr_j 为用户 j 的实际评分; M 为用户数; Δr 为最大值 r_{\max} 与最小值 r_{\min} 差的绝对值。

Precision 定义为一个用户对已推荐产品感兴趣的可能性,是用户推荐列表的数量与总的产品数之比。计算公式为:

$$\text{precision} = \frac{1}{N} \sum_{j=1}^N \frac{|T_j \cap \text{Rt}_j|}{|T_j|} \quad (7)$$

其中, N 为用户数; T_j 为用户 j 排名; Rt_j 为用户 j 选取的物品评分。

2.2 实验结果与分析

在个性化推荐系统中,有三种其他的推荐方法,包括基于物品(Items-Based)的协同过滤推荐方法、基于评分(Rates-Based)的协同过滤推荐方法以及模糊偏好集的多属性决策协同过滤推荐方法(F-MADM)。基于物品的协同过滤方法的目的是找到用户选择相似物品的邻居,通过余弦相似度方法,并传统协同过滤算法相结合产生推荐结果。基于评分的方法通过计算余弦相似度得到已评分用户的邻居用户,再绑定传统的协同过滤算法产生推荐结果。F-MADM 方法是基于模糊用户偏好集构建用户偏好模型,然后绑定协同过滤算法产生推荐结果。

更低的 NMAE 是具有更好精确度的推荐。从图 2 中可以发现,文中方法的 NMAE 最低, F-MADM 的 NMAE 次之,基于物品和基于评分的方法的 NMAE 最不好,但是两者相近。相比而言,文中推荐方法的精确性较好。

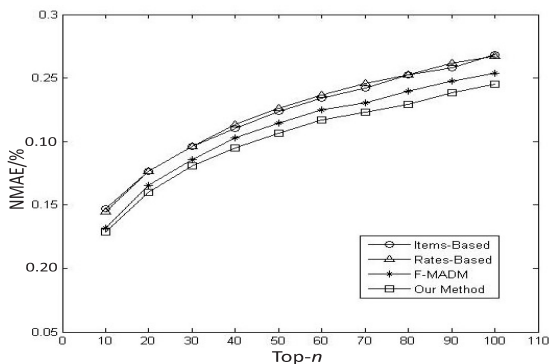


图 2 归一化平均绝对差对比

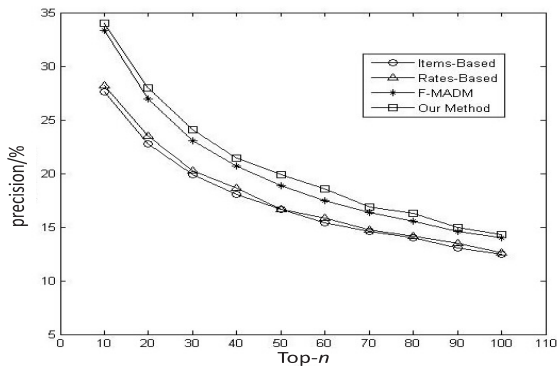


图 3 精确度对比

如图 3 所示,当横坐标为 60 时,文中方法与 F-MADM 相比,精确度提高超过 6%。同时也可以发现,基于物品的推荐方法的性能与基于评分的性能是接近的。与以上两种方法相比,可以发现在横坐标为 60

时,精确度提高超过 20%,表明文中方法的精确性较好。

3 结束语

针对个性化推荐技术存在的问题,提出一种新的推荐方法。该方法能实现自动提取物品的领域特征值,并通过领域关键词验证过滤冗余的领域特征值,由得到的领域特征值集构建用户偏好模型,并将用户偏好模型与协同过滤算法绑定产生推荐结果。与相应的推荐方法进行了对比实验,结果表明,文中方法具有更好的准确性,同时能够有效地挖掘领域特征值并构建用户偏好模型。未来研究将会增加领域特征值的数量以提高推荐方法的多样性,同时在其他领域验证该方法的准确性。

参考文献:

- [1] Huang H, Huang J, Zivarras S G, et al. A personalized recommendation algorithm based on Hadoop[C]//5th international conference on electronics information and emergency communication. [s. l.]: IEEE, 2015: 406–409.
- [2] Pera M S, Ng Y K. Analyzing book-related features to recommend books for emergent readers[C]//Proceedings of the 26th ACM conference on hypertext & social media. [s. l.]: ACM, 2015: 221–230.
- [3] Chen T, Han W L, Wang H D, et al. Content recommendation system based on private dynamic user profile[C]//International conference on machine learning and cybernetics. [s. l.]: IEEE, 2007: 2112–2118.
- [4] 王全民, 王 莉, 曹建奇. 基于评论挖掘的改进的协同过滤推荐算法[J]. 计算机技术与发展, 2015, 25(10): 24–28.
- [5] 黄创光, 印 鉴, 汪 静, 等. 不确定近邻的协同过滤推荐算法[J]. 计算机学报, 2010, 33(8): 1369–1377.
- [6] 李 聪, 梁昌勇, 马 丽. 基于领域最近邻的协同过滤推荐算法[J]. 计算机研究与发展, 2008, 45(9): 1532–1538.
- [7] 郑 丹, 王名扬, 陈广胜. 基于 Weighted-slope One 的用户聚类推荐算法研究[J]. 计算机技术与发展, 2016, 26(4): 51–55.
- [8] 高 倩, 何聚厚. 改进的面向数据稀疏的协同过滤推荐算法[J]. 计算机技术与发展, 2016, 26(3): 63–66.
- [9] Hu L, Song G, Xie Z, et al. Personalized recommendation algorithm based on preference features[J]. 清华大学学报: 自然科学英文版, 2014, 19(3): 293–299.
- [10] Jung S Y, Hong J H, Kim T S. A statistical model for user preference[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6): 834–843.
- [11] Choi S H, Jeong Y S, Jeong M K. A hybrid recommendation method with reduced data for large-scale application[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2010, 40(5): 557–566.
- [12] Ha V, Haddawy P. Similarity of personal preferences: theoretical foundations and empirical analysis[J]. Artificial Intelligence, 2003, 146(2): 149–173.
- [13] Chen Y Y, Cheng A J, Hsu W H. Travel recommendation by mining people attributes and travel group types from community-contributed photos[J]. IEEE Transactions on Multimedia, 2013, 15(6): 1283–1295.
- [14] Zhang J, Peng Q, Sun S, et al. Employing F-MADM to derive user preference model from item features and rating information for personalized recommendation[C]//IEEE international conference on information and automation. [s. l.]: IEEE, 2015: 2997–3002.
- [15] Tewari A S, Kumar A, Barman A G. Book recommendation system based on combine features of content based filtering, collaborative filtering and association rule mining[C]//International conference on advance computing. [s. l.]: IEEE, 2014: 500–503.
- [16] Chen J H, Chao K M, Shah N. Hybrid recommendation system for tourism[C]//10th international conference on e-business engineering. [s. l.]: IEEE, 2013: 156–161.
- [17] Chen R C, Huang Y H, Bau C T, et al. A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection[J]. Expert Systems with Applications, 2012, 39(4): 3995–4006.