文章编号：2095-2163(2020)02-0071-04 中图分类号：TP391.3 文献标志码：A

网站用户行为分析及服务推荐研究

张婉婷，赵 敏
（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘 要：本文运用基于物品的协同过滤推荐算法设计法律资讯信息推荐系统。介绍了该算法需要的数学理论知识，如相似度计算方法、系统评估方法和 KNN 算法等。详细解释了基于物品的协同过滤推荐算法及其具体实现步骤。最后构建法律资讯信息推荐系统并对系统做出评估。通过实验仿真分析，发现基于物品的协同过滤算法在物品种类丰富，用户个性化需求强烈的领域优势明显。其相关的推荐和解释利用用户的历史行为数据，结果让用户信服。

关键词：协同过滤；相似度；KNN；法律资讯信息推荐系统

Website user behavior analysis and service recommendation research

ZHANG Wanting, ZHAO Min
(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

【Abstract】 This paper uses collaborative filtering recommendation algorithm based on items to design legal information recommendation system. Firstly, the paper introduces the theoretical knowledge of the algorithm, such as similarity calculation method, system evaluation method and KNN algorithm. Then, the paper explains the collaborative filtering recommendation algorithm based on items and its implementation steps in detail. Finally, the paper constructs a legal information recommendation system and evaluate the system. Through the experiment analysis, it is found that the collaborative filtering algorithm based on goods has obvious advantages in the fields of rich kinds of goods and strong personalized needs of users. Its relevant recommendation and interpretation make use of the historical behavior data of users, and the results are convinced.

【Key words】 collaborative filtering; similarity; KNN; legal information recommendation system

0 引 言

随着互联网的广泛应用和迅猛发展，海量的信息也随即涌现，人们已经进入信息过载的时代[1]。对于想要获取信息的用户，靠自己从庞大的数据中获取需要的信息则存在相当难度。此外，对于提供信息的一方，如何让自己提供的信息受到用户的关注，也并非易事。为此有众多高校、研究机构积极开展此项研究[2]。其中，推荐系统就是解决该问题的一项重要工具。推荐系统利用数据挖掘等相关的数据处理技术，能根据用户的历史行为数据，来自动生成用户可能感兴趣的信息，并将其推荐给用户，最终为用户提供个性化信息推荐服务[3]。

1 国内外推荐系统研究现状

1995 年，美国人工智能协会上首次提出了个性化推荐系统。随后各大高校和研究部门开始对个性化推荐系统进行深入研究。与此同时，又陆续推出了多种推荐算法，其中，目前获得较大成功、有着较为广泛应用的则是基于协同过滤的推荐算法[4]。

针对协同过滤推荐算法自身的问题（冷启动和稀疏数据问题等）[5]，学界提出了多种改进的方法。对于冷启动问题，文献[6]中给出了一种基于新用户的冷启动推荐方法。将用户模型和信任/不信任网络结合，以此来识别值得信任的用户，取得了较好的效果。对于稀疏数据问题，文献[7]将贝叶斯网络引入到协同过滤当中，并通过对网络的进一步优化，提高了系统的准确率[8]。对于协同过滤推荐算法存在的其他问题，韩林峄等人[9]基于用户之间的共同/不同特征，创建了新的相似度计算方法，并在基于 K 近邻的协同过滤算法中得到应用。

在协同过滤推荐算法基础上，文献[10]改进了协同过滤推荐算法，并比传统方法的推荐准确度更高。即使用户评分数据极端稀疏，推荐效果仍较好。另外，为了满足不同用户的多样化需求，产生了多种混合的推荐算法。Girardi 等人[11]通过将领域本体技术与协同过滤进行混合，并取得了理想的推荐效果。Vekariya 等人[12]通过将知识和协同过滤相结合构造混合推荐算法，不仅证明了该算法的有效性，而且取得了较好的效果。

作者简介：张婉婷(1993-)，女，硕士研究生，主要研究方向：数据处理、模型预测控制；赵 敏(1979-)，女，博士，讲师，硕士生导师，主要研究方向：预测控制。

收稿日期：2019-11-12

在前人的思想上,本文主要通过基于物品的协同过滤推荐算法对法律资讯平台建立个性化推荐系统。一个完整且完善的推荐系统通常具有用户行为信息收集、用户喜好模型建立和信息推荐等功能。

2 推荐算法

2.1 相似度计算

在推荐系统中,最重要的计算就是用户-物品偏好二维矩阵运算。可以将某个用户对所有物品的偏好作为一个向量,计算用户之间的相似度,或将所有用户对某个物品的偏好作为一个向量,计算物品之间的相似度[13]。常用的相似度计算方法有:余弦相似度法、欧几里德距离法和皮尔逊相关系数法。基于优化思想的考虑,本文的法律资讯信息推荐系统采用皮尔逊相关系数法。

皮尔逊相关系数可表示两变量间的线性相关程度。该系数取值-1 到+1,越接近 1 或-1 说明变量之间线性关系越强。如果一个变量增大,另一个变量也增大,则变量之间正相关,相关系数大于 0。如果一个变量增大,另一个变量却减小,则变量之间负相关,相关系数小于 0。如果相关系数等于 0,则表明变量之间不存在线性相关关系[14]。

皮尔逊相关系数计算公式为:

$$sim(i,j)=\frac{\sum_{d\in I,i,j}(R_{i,d}-\overline{R_i})\,(R_{j,d}-\overline{R_j})}{\sqrt{\sum_{d\in I,i,j}(R_{i,d}-\overline{R_i})^2}\ \sqrt{\sum_{d\in I,i,j}(R_{j,d}-\overline{R_j})^2}}. \tag{1}$$

其中,$R_{i,d}$ 表示用户 i 对物品 d 的评分,$\overline{R_i}$、$\overline{R_j}$ 表示用户 i 和用户 j 对所打分物品的平均评分[15]。

2.2 系统评估

本文法律资讯信息推荐系统主要采用均方根误差 (*RMSE*)、平均绝对误差(*MAE*) 对系统进行评估,用来评估本文构建的推荐系统的推荐质量。对此可做研究分述如下。

(1)均方根误差 (*RMSE*)。 可通过式(2)进行计算:

$$RMSE=\sqrt{\frac{\sum_{u,d\in T}(r_{ud}-\hat{r}_{ud})^2}{|T|}}, \tag{2}$$

(2)平均绝对误差 (*MAE*)。 可通过式(3)进行计算:

$$MAE=\frac{\sum_{u,d\in T}|r_{ud}-\hat{r}_{ud}|}{|T|}. \tag{3}$$

其中,T 表示测试的数据集,r_{ud}、$\hat{r}_{ud}$ 表示用户 u 对物品 d 的预测评分和实际评分。

2.3 K-近邻算法

K-近邻算法(KNN)的原理可表述为:特征空间中有 k 个最相邻的样本,当某一个不知类别的样本,与样本中的大多数属于同一个类别时,就认为该样本也属于这个类别,且该样本也具有该类别的其它特征。K-近邻算法在分类时,用与未知类别的样本最邻近的一个(或几个)样本的类别,来决定待分类样本所属的类别。

这里,给出该算法的执行步骤详述为:

(1)计算已知类别的样本点与待分类样本点之间的距离。

(2)对计算出的距离进行排序。

(3)确定与待分类样本点距离最小的数个已知类别的样本点。

(4)计算距离最小的数个样本点所在类别的出现频率。

(5)返回距离最小的数个样本点中出现频率最高的类别作为待分类样本点的预测分类。

3 协同过滤

3.1 协同过滤的选择

协同过滤主要依靠用户历史行为数据和属性的相近性来实现个性化推荐服务。协同过滤不仅可以对喜好相近的用户进行信息收集,还可以将有用的信息推送给其他相似用户作为参考。常用的协同过滤方法有基于用户(user-based)与基于物品(item-based)的协同过滤。本文是以某大型法律资讯网站提供的数据为基础来构建推荐系统。

基于用户的协同过滤推荐方法是通过和目标用户相似的用户的兴趣喜好来预测目标用户的兴趣偏好,并推荐目标用户可能喜欢的物品[16]。该方法是以用户访问行为的相似性为基础,来推荐用户可能感兴趣的内容和资源。依此建立的法律资讯信息推荐系统的基本思想是将与自己有相同或相似兴趣偏好的用户所喜欢的法律知识(网站)推荐给自己。

与基于用户的推荐方法相反,基于物品的协同过滤推荐方法是根据用户平时浏览过的一些法律知识(网站)及其喜好程度,通过计算各不同法律知识(网站)之间的线性相关程度,推荐给用户与之前访问法律知识(网站)相似的其它法律知识(网站)。

随着互联网的快速发展,网站用户的数目也越来越庞大。考虑到基于用户与基于物品的协同过滤的区别,在实际应用中,计算用户的兴趣相似度矩阵越来越困难(例如庞大的稀疏矩阵)。运算的时间

复杂度和空间时间复杂度呈指数增长。基于用户的协同过滤方法不利于实际中的研究应用，而且对推荐的结果也不能做出很好的解释[17]。因此本文法律资讯信息推荐系统是在基于物品的协同过滤前提下进行研发的，主要是对法律咨询中婚姻这一大类进行小类别的推荐。

3.2　基于物品的协同过滤

该算法可以理解为，法律知识 A 和法律知识 B 具有很大的相似度，那么喜欢法律知识 A 的用户也大都喜欢法律知识 B。

研究可知，法律资讯信息推荐系统中基于物品的协同过滤推荐思路如图 1 所示。由图 1 可知，此方法的基本思想是把与用户喜欢的法律知识(网站)相似度较高的其它法律知识(网站)推荐给用户。

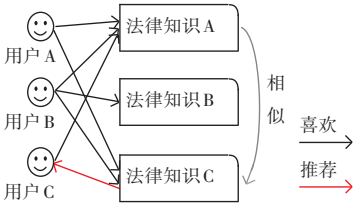

图 1　法律资讯信息推荐系统中基于物品的协同过滤推荐思想

Fig. 1　Item-based collaborative filtering recommendation in legal information recommendation system

基于物品的协同过滤推荐算法首先会收集用户、物品信息。建立用户兴趣模型。计算目标物品与其他物品的相似度，并对相似度排序。然后根据推荐算法给用户推荐和目标物品最相似的 i 个物品[18]。基于物品的协同过滤推荐算法中用户 u 对物品 n 的感兴趣程度的度量公式见如下：

$$p(u,n)=\frac{\sum(q_{Nn}*d_{uN})}{\sum|q_{Nn}|}. \tag{4}$$

其中，q_{Nn} 表示物品 n 和目标物品 N 的相似度，d_{uN} 表示用户 u 对目标物品 N 的感兴趣程度。

基于物品的协同过滤推荐算法的设计步骤可以描述为：

(1)计算目标物品与其他物品之间的相似度。

(2)根据用户的历史行为、物品间相似度的大小等生成推荐列表。

4　设计法律资讯信息推荐系统

4.1　操作步骤

(1)数据预处理。数据预处理主要是对得到的用户原始的浏览信息进行一定的处理，例如筛选、去除无用的信息等操作。本文采用的是某大型法律资讯网站提供的数据，每一个用户数据包括 realId、fullURL、网页标题、网站所属法律知识的类别等。本次预处理操作包括筛选出访问婚姻相关信息的用户数据、删除有残缺的数据、把具有翻页功能的网站还原到初始网页等。

(2)用户偏好程度设置。考虑到无法真实地获知用户对哪种法律知识比较关注(喜好)，本文通过统计用户对某一类法律知识网页访问的(不同)次数来作为用户的关注(喜好)程度，并做归一化处理。

(3)确定输入数据模型。本次系统的建立是采用 Python 的 surprise 包，因此输入数据模型将根据其内部的要求采用如下数据模型，即：<用户 id><物品 id><用户偏好得分>。

(4)相似度矩阵计算。本文采用基于 KNN 算法的协同过滤推荐系统，采用皮尔逊相关系数计算其相似度。

(5)得出推荐结果。根据某一用户之前的访问状况给出某一类别法律知识，推荐给用户可能关注的不同法律知识及网站。

4.2　运行推荐系统

法律资讯信息推荐系统以某大型法律资讯网站提供的数据为基础。对法律知识中的婚姻资讯进行推荐系统的建立和实验验证。

以婚姻中“财产分割”为例，推荐给关注该法律知识的用户 5 个相关的法律知识类别并列出相关的网站。

(1)法律资讯信息推荐系统推荐结果。见图 2。

图 2　法律资讯信息推荐系统的推荐结果

Fig. 2　Recommended results of the legal information recommendation system

由图 2 可知，法律资讯信息推荐系统给出了与婚姻中“财产分割”相似度最高的五类法律知识及相关网站，即：“继承法”、“抚养费”、“离婚诉讼”、

“非婚生子女”和“哺乳期”。综上内容与“财产分割”相关性很高。

(2)法律资讯信息推荐系统评估参数。见图 3。

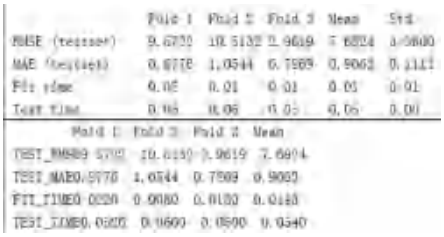

图 3 法律资讯信息推荐系统的评估参数

Fig. 3 Evaluation parameters of the legal information recommendation system

(3)法律资讯信息推荐系统评估结果。研究评估结果如图 4 所示。

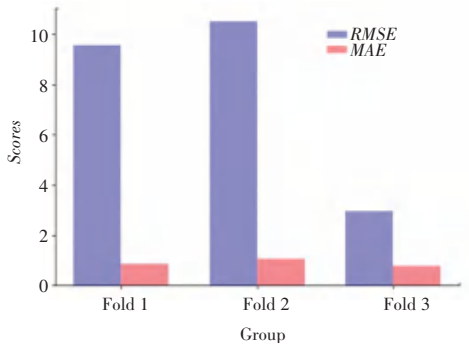

图 4 法律资讯信息推荐系统的评估结果

Fig. 4 Evaluation results of the legal information recommendation system

由图 2~图 4 可知,法律资讯信息推荐系统为关注婚姻中“财产分割”的用户推荐的结果,比较符合实际情况。此外,该系统在 *RMSE* 和 *MAE* 上也有较好的表现。

5 结束语

本文的法律资讯信息推荐系统是基于物品的协同过滤设计的个性化推荐系统。文中通过一系列设计研发步骤,最终取得了令用户信服的仿真测试结果。

在整个算法研究过程中,考虑到用户信息数据的庞大[19],本文只对法律知识中婚姻资讯进行推荐系统的建立和实验验证。虽然如此,系统运行速度仍然较慢,还要不断改进算法,尽量使其运行得更快。此外,算法本身的数据稀疏问题、用户兴趣迁移问题等,也影响着推荐效果,这些均有待更进一步的研究去改进与完善。

参考文献

[1] 赵耀培. 新媒介环境下主流媒体的经营创新之道[J]. 新媒体研究,2019,5(4):59.
[2] 金石. 基于运营商管道大数据的智能电商推荐系统[D]. 南京:南京邮电大学,2018.
[3] 赵伟明. 基于用户行为分析和混合推荐策略的个性化推荐方法研究[D]. 北京:北京工业大学,2014.
[4] 王强. 基于协同过滤的个性化推荐算法研究及系统实现[D]. 成都:西南交通大学,2017.
[5] 李晓菊. 协同过滤推荐系统中的数据稀疏性及冷启动问题研究[D]. 上海:华东师范大学,2018.
[6] CHEN C C, WAN Yuhao, CHUNG M C, et al. An effective recommendation method for cold start new users using trust and distrust networks[J]. Information Sciences,2013,224:19.
[7] SU Xiaoyuan, KHOSHGOFTAAR T M. Collaborative filtering for multi-class data using belief nets algorithms[C]//18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI '06). Arlington, VA, USA:IEEE,2006:497.
[8] 郭昱锦. 面向 O2O 用户行为分析的个性化推荐算法研究与应用[D]. 保定:河北大学,2017.
[9] 韩林峄,吴晟. 通过评分特征优化基于 K 近邻的协同过滤算法[J]. 信息技术,2018,42(12):75.
[10] 赵丽嫚. 一种新型的协同过滤推荐算法[D]. 南京:南京邮电大学,2013.
[11] GIRARDI R, MARINHO L B. A domain model of Web recommender systems based on usage mining and collaborative filtering[J]. Requirements Engineering,2006,12(1):23.
[12] VEKARIYA V, KULKARNI G R. Hybrid recommender systems: Content-boosted collaborative filtering for improved recommendations[C]//Proceedings of the 2012 International Conference on Communication Systems and Network Technologies. USA:IEEE,2012: 649.
[13] 阳志梁. 基于新型信任模型的自适应推荐方法研究—以内蒙古自治区为例[D]. 上海:上海师范大学,2018.
[14] 李舒婷. NDVI 时空变化及其与牛羊肉产量关系分析—以内蒙古自治区为例[D]. 北京:中国科学院大学(中国科学院遥感与数字地球研究所),2018.
[15] 马宏伟,张光卫,李鹏. 协同过滤推荐算法综述[J]. 小型微型计算机系统,2009,30(7):1282.
[16] 张海蛟. 一种协同过滤中相似度计算和近邻用户查找算法研究[D]. 长沙:湖南大学,2017.
[17] 陈阿龙. 推荐系统用户冷启动问题相关研究[D]. 长沙:国防科学技术大学,2016.
[18] 谭立云,刘琳,苏鹏. 图书借阅推荐系统算法的 Python 实现[J]. 科学技术创新,2018(22):84.
[19] 毛宜钰. 协同过滤推荐算法的稀疏性与可扩展性问题研究[D]. 湘潭:湖南科技大学,2017.

(上接第 70 页)

[6] 刘川,刘景林. 基于 Simulink 仿真的步进电机闭环控制系统分析[J]. 测控技术,2009,28(1):44.
[7] 郭豪,李宝慧,赵树忠. 基于模糊 PID 的步进电机建模与仿真[J]. 机械工程与自动化,2018(2):167.
[8] 邱宏超,刘教瑜,肖杰,等. 三相混合式步进电机的矢量控制的研究[J]. 工业控制计算机,2016,29(5):140.
[9] 王悦. 基于模糊 PI 控制方法的三相混合式步进电机驱动器设计[D]. 杭州:浙江工业大学,2013.
[10] 唐志航,俞立. 模糊控制参数的设计[J]. 电气时代,2002(11):75.