文章编号:1000-5862(2016)02-0145-04

基于决策树分类的个性化农产品移动信息服务系统

陈亚慧,叶继华*
(江西师范大学计算机信息工程学院,江西 南昌 330022)

摘要:针对农产品移动信息服务的需求,结合分类算法和个性化推荐算法,提出了一种基于分类的推荐算法.利用决策树分类方法对农产品进行分类,获得分类后的数据,采用协同过滤算法分析分类数据,查找兴趣相似的用户,将感兴趣的农产品信息推荐给正在使用系统的用户.实验结果表明:与传统的推荐方法及相比,该系统向用户推荐了兴趣度更高的农产品移动信息.

关键词:决策树算法;个性化推荐;协同过滤算法;农产品移动信息服务
中图分类号:TP 391 文献标志码:A DOI:10.16357/j.cnki.issn1000-5862.2016.02.07

0 引言

随着 3G 和 4G 网络时代的到来以及移动设备在农村的普及,利用移动互联网推广农产品成为一种有效的手段.建立以移动互联网为平台的农产品信息服务系统能够有效地解决农业信息化问题,促进农产品信息高效畅通,增加农产品的销售渠道,同时为农业管理部门提供信息服务.

推荐系统有 3 个重要的模块:用户建模模块、推荐对象建模模块、推荐算法模块[1].用户模型对用户进行分类和识别,帮助系统更好地理解用户特征和类别,理解用户的需求和任务,从而更好地实现用户所需要的功能.推荐对象的描述主要有基于内容的方法和基于分类的方法两大类方法.基本推荐策略包括:基于内容的推荐[2]、协同过滤推荐[3-4]、基于网络结构[5-6]的推荐等.基于内容的推荐受到推荐对象特征提取能力的限制较为严重,并且较难为用户发现新的感兴趣的信息,还存在冷启动问题、数据量较大、不同语言的描述的用户模型和推荐对象模型无法兼容等问题.协同过滤存在冷启动问题、稀疏性问题,系统开始时推荐质量差及推荐质量取决于历史数据集.

各种推荐方法有各自的优缺点,在实际应用中可以针对具体问题采用推荐策略的组合进行推荐,即所谓的组合推荐,主要混合思路有 2 种[7-8]:推荐结果混合和推荐算法的混合.本文结合决策树分类方法和协同过滤算法,提出了一种基于分类的推荐算法,并应用于农产品的推荐,结合农产品信息的处理结果表明了本方法有一定的改进效果.

1 算法分析

本文利用决策树分类方法对农产品进行分类,获得分类后的数据,采用协同过滤算法分析分类数据,查找兴趣相似的用户,将其感兴趣的农产品信息推荐给正在使用系统的用户.

1.1 决策树分类

决策树分类方法的基本思想[9-10]是利用数据库里面的数据自动地构造决策树,然后根据该决策树获得分类的数据.数据来源于后台的数据库中,农产品信息中基本包含了农产品名称(name)、农产品种类(type)以及农产品产地(area)等字段信息,根据这些字段利用决策树方法对后台农产品数据进行分类,步骤如图 1 所示.(i)从树的根节点处的所有数据开始(农产品供应信息表所有数据),对每一条数据选择一个属性.对属性的每一个值产生一分枝.分枝属性值对应的数据被移到新生产的子节点上.(ii)将这个方法递归地应用于每个子节点,直到一个节点上的所有数据都分区到某个类中.(iii)到达决策树的叶节点的每条路径表示一个分类规则.

1.2 协同过滤个性化推荐

协同过滤个性化推荐方法的基本思想[1]就是

收稿日期:2015-10-27
基金项目:国家自然科学基金(61462042)和江西师范大学研究生创新基金(YJS2013082)资助项目.
通信作者:叶继华(1966-),男,江西上饶人,教授,主要从事物联网和图像处理的研究.

分析用户兴趣,在用户群中找到指定用户的相似(兴趣)用户,综合这些相似用户对某一信息的评价,形成系统对该指定用户对此信息的喜好程度预测,然后进一步将用户感兴趣的信息推荐给用户,具体实现如图 2 所示.

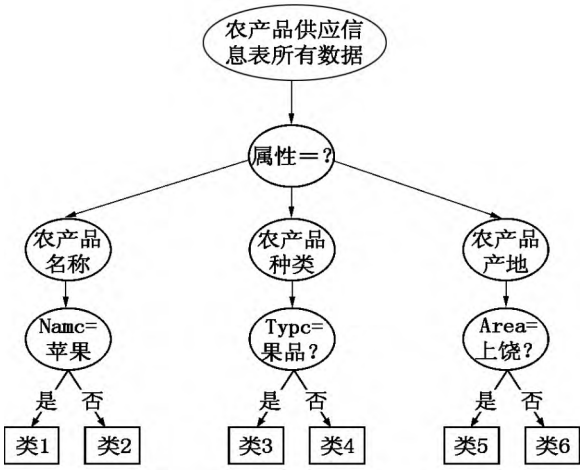

图 1 决策分类

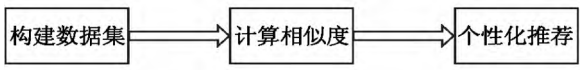

图 2 协同过滤推荐过程

1.2.1 **构建数据集** 数据集是协同过滤算法的关键. 首先,根据分类的农产品供应信息可知,存在用户 m 个,项目 n 个,记录 t 个. 然后对数据集进行描述: 实验所用的数据集就是 m 个系统用户对 n 个农产品(项目)的 t 个真实调用所产生的访问记录(用户访问农产品信息). 最后,构建一个用户-项目访问矩阵作为协同过滤算法的输入,如表 1 所示. R_{ij} 表示第 i 个用户对第 j 个项目的访问情况(R_{ij} 的取值为 0 或者 1,表示用户是否访问该项目,0 表示没有访问记录,1 表示存在访问记录).

1.2.2 **计算相似度** 协同过滤算法的主要思想是通过查找与目标用户相似的近邻用户,通过近邻用户的访问记录对目标用户产生推荐. 本文通过使用皮尔逊相关系数(Pearson Correlation Similarity)来计算相似度. 假设用户 i 和用户 j 共同访问的项目集合为 I_{ij},则利用皮尔逊相关可得到两者的相似性 $Sim(i,j)$ 为

$$Sim(i,j) = \frac{\sum_{k\in I_{ij}}(R_{ik}-\overline{R_i})(R_{jk}-\overline{R_j})}{\sqrt{\sum_{k\in I_{ij}}(R_{ik}-\overline{R_i})^2}\sqrt{\sum_{k\in I_{ij}}(R_{jk}-\overline{R_j})^2}}, \quad (1)$$

其中 R_{ik} 和 R_{jk} 分别代表用户 i 和用户 j 对项目 k 的访问,$\overline{R_i}$ 和 $\overline{R_j}$ 分别代表用户 i 和用户 j 在所有项目的评分平均值. 相似性 $Sim(i,j)$ 值越大,表示这两者的相似性越大,反之,则相似性越小.

表 1 用户-项目访问矩阵

	项目 1	项目 2	项目 3	…	项目 n
用户 1	R_{11}	R_{12}	R_{13}	…	R_{1n}
用户 2	R_{21}	R_{22}	R_{23}	…	R_{2n}
用户 3	R_{31}	R_{32}	R_{33}	…	
…	…	…	…	…	…
用户 m	R_{m1}	R_{m2}	R_{m3}	…	R_{mn}

通过相似性度量方法计算完用户间的近邻关系之后,根据相似性排序从大到小依次选择前面的 K 个最相似的用户作为目标用户的近邻集合.

1.2.3 **个性化推荐** 通过计算得到目标用户的近邻集合,再根据近邻集合中用户的评分来预测目标用户对于未评分项目的评分. 本文采用的预测方法为

$$P_{ik} = \overline{R_i} + \frac{\sum_{j\in KNB} Sim(i,j) \times (R_{jk}-\overline{R_j})}{\sum_{j\in KNB} Sim(i,j)}, \quad (2)$$

P_{ik} 代表用户 i 对项目 k 的预测评分,KNB 表示目标用户 i 的近邻集合,$\overline{R_i}$ 和 $\overline{R_j}$ 分别代表用户 i 和用户 j 在所有已评分项目上的平均评分,$Sim(i,j)$ 表示用户 i 和用户 j 的相似度,R_{jk} 表示用户 j 对项目 k 的评分值. 项目的预测评分值不为 0 说明用户对该项目存在兴趣,可以推荐给用户.

2 实验结果及分析

2.1 实验平台的构建

基于决策树分类的个性化农产品信息服务系统采用了客户端/服务器(B/S)结构,采用 Java 技术,基于 Eclipse + MyEclipse + MySQL + Android 软件开发环境.

2.2 实验结果及分析

2.2.1 **实验 1 基于决策树的分类** 假设表 1 是一个农产品情况的数据,描述农产品的主要特征属性有: 名称、种类、产地,而这些特征属性可能取值为: 名称 = {南瓜,丝瓜,菠菜,苹果,西瓜,脐橙,大豆,芝麻,玉米,紫罗兰,向日葵,含羞草},种类 = {蔬菜,果品,粮食,花卉},产地 = {南昌,九江,景德镇,萍乡,新余,鹰潭,赣州,宜春,上饶,吉安,抚州}. 则农产品的简单描述为: {名称: 菠菜,种类: 蔬菜,产地: 南昌},获得的农产品信息分类如表 2 所示.

由表 2 中的数据可知,总共有 14 个项目对象,若以不同属性进行划分,则获得不同的类别.

表 2　农产品信息分类

序号	属性			分类结果			
	农产品名称	农产品种类	农产品产地	以名称划分类别	以种类划分类别	以产地划分类别	以 3 个属性综合划分类别
1	南瓜	蔬菜	南昌	1	1	1	1
2	菠菜	蔬菜	九江	2	1	2	2
3	苹果	果品	宜春	3	2	3	3
4	玉米	粮食	鹰潭	4	3	4	4
5	向日葵	花卉	九江	5	4	2	5
6	芝麻	粮食	上饶	6	3	5	6
7	脐橙	果品	新余	7	2	6	7
8	紫罗兰	花卉	赣州	8	4	7	8
9	丝瓜	蔬菜	九江	9	1	2	9
10	西瓜	果品	萍乡	10	2	8	10
11	菠菜	蔬菜	上饶	2	1	5	11
12	含羞草	花卉	抚州	11	4	9	12
13	苹果	果品	鹰潭	3	2	4	13
14	芝麻	粮食	上饶	6	3	5	14

2.2.2　**实验 2　基于不同分类、不同近邻参数的个性化推荐**　评价推荐系统性能的好坏通常用推荐的精确度和推荐效率 2 个指标进行衡量. 精确度的衡量最典型的指标[11-12]是平均绝对误差(Mean Absolute Error，*MAE*)和平均平方误差(Mean Squared Error，*MSE*)以及标准平均误差(Normalized Mean Absolute Error，*NMAE*). 它们的计算形式分别为

$$MAE = \frac{1}{n}\sum_{a=1}^{n} |p_{ia} - r_{ia}|, \tag{3}$$

$$MSE = \sqrt{\frac{1}{n}\sum |p_{ia} - r_{ia}|^2}, \tag{4}$$

$$NMAE = MAE/(r_{\max} - r_{\min}), \tag{5}$$

其中 n 为系统中用户 i 打分产品的个数，P_{ia} 和 r_{ia} 分别为预测打分和实际打分. n_i 为系统中用户-产品对的个数. $r_{\min}$ 和 $r_{\max}$ 分别为用户打分区间的最小值和最大值. 本文选用 *MAE* 指标进行比较，*MAE* 值是一种推荐质量度量方法，值越小，则推荐效果越好.

以实验 1 的项目对象为基础，假设存在 100 个用户对这 14 个项目对象存在 850 个访问记录，每个用户至少有 7 个访问记录，推荐过程中存在近邻参数的变化，且最近邻参数的变化对系统推荐的效果有影响，则在近邻参数变化下，系统推荐效果如图 3 所示.

根据图 3 中的数据可知，随着近邻参数的变化，不同方法的平均绝对偏差 *MAE* 的值逐渐减小，说明近邻参数变化对个性化推荐的效果有着一定的影响. 从图 3 可以看出，分类后进行个性化推荐的 *MAE* 值小于传统的个性化推荐，说明在同等影响条件下，基于分类的个性化推荐效果要好于传统的个性化推荐，然而，综合分类的个性化推荐效果和传统的个性化推荐效果是一样的，这是由于实验中综合分类的项目对象划分结果和原始的项目对象是一样的.

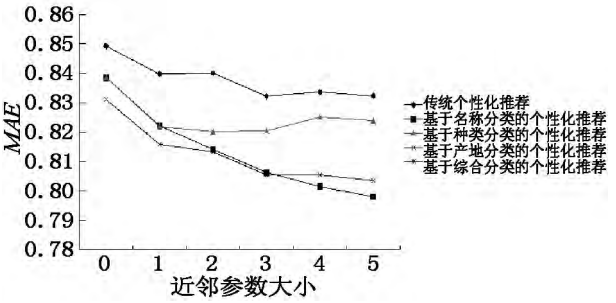

图 3　不同近邻参数大小推荐效果的影响

2.2.3　**实验 3　不同推荐算法的比较**　以实验 1 的项目对象为基础，假设存在 100 个用户对这 14 个项目对象存在 1 650 条访问记录，每个用户至少有 14 个访问记录，推荐过程中存在近邻参数的变化，且最近邻参数的变化对系统推荐的效果有影响，则在近邻参数变化下，原始推荐方法、基于分类的推荐方法以及用户间多相似度协同过滤推荐算法[13]推荐效果如图 4 和表 3 所示.

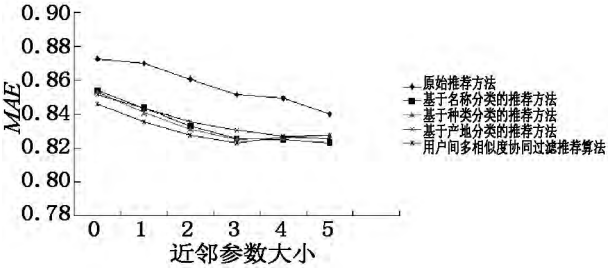

图 4　不同推荐方法的推荐效果

由图 4 可知，在不同的近邻参数下，基于分类的推荐方法获得 *MAE* 值均小于原始推荐方法的 *MAE* 值. 用户间多相似度协同过滤推荐算法随着近邻参数的变化具有不同的推荐效果，近邻参数在 0 ~ 3 的变化范围内，*MAE* 值是最小的，推荐效果最好，近邻参数在 4 ~ 5 范围内变化时，相比本文的推荐方法获

得的 *MAE* 值偏大,且由表 3 可知,用户间多相似度协同过滤推荐算法推荐过程中使用的时间相比基于分类的推荐方法偏长,影响了系统的实时性. 因此,综合考虑推荐效果及推荐时间,基于分类的推荐方法更适用于实际应用中的农产品移动信息服务系统.

表 3 不同推荐方法的推荐时间 s

推荐方法	最近邻参数					
	0	1	2	3	4	5
传统推荐	1.471	1.465	1.464	1.468	1.469	1.462
基于名称分类的推荐	1.324	1.317	1.302	1.285	1.231	1.247
基于种类分类的推荐	1.268	1.264	1.253	1.235	1.227	1.216
基于产地分类的推荐	1.302	1.297	1.292	1.276	1.269	1.243
文献 8 中的多相似度推荐	1.431	1.426	1.421	1.419	1.414	1.413

3 结束语

本文据现有农产品信息服务系统的不足,提出将分类算法结合个性化推荐算法实现个性化农产品移动信息服务系统. 这不仅能够建立一个农产品信息高效流通体系,避免了农产品运输过程中的高损失率和高成本,实现农民、运营商以及零售商等多方共赢,还能促进农业信息化的发展.

4 参考文献

[1] 王国霞,刘贺平. 个性化推荐系统综述 [J]. 计算机工程与应用,2012,48(7):66-76.

[2] 刘玮. 电子商务系统中的信息推荐方法研究 [J]. 情报科学,2006,24(2):300-303.

[3] Sarwar B,Karypis G,Konstan J,et al. Item-based collaborative filtering recommendation algorithms [C]. New York: ACM Press,2001:285-295.

[4] Karypis G. Evaluation of item-based top-n recommendation algorithms [C]//Proceedings of the 10th International Conference on Information and Knowledge Management,2001.

[5] Zhou Tao,Ren Jie,Medo M,et al. Bitpartite network projection and personal recommendation [J]. Phys Rev E,2007.76(4):70 - 80.

[6] Zhou Tao,Jing Luoluo,Su Riqi,et al. Effect of initial configuration on network-based recommendation [J]. Europhys Lett,2008,81(5):15-18.

[7] Basu C,Hirsh H,Cohen W. Recommendation as classfication: using social and content-based information in recommendation [C]. Mello Park: AAAI Press,1998:714-720.

[8] Claypool M,Gokhale A,Miranda T,et al. Combining content-based and collaborative filters in an online newspaper [C]. New York: ACM Press,1999.

[9] 刘小虎,李生. 决策树的优化算法 [J]. 软件学报,1998,9(10):797-799.

[10] Erick T Byrd, Larry Gustke. Using decision trees to identify tourism stakeholders [J]. Journal of Place Management and Development,2011,4(2):148-168.

[11] 许海玲. 互联网推荐系统比较研究 [J]. 软件学报,2009,20(2):350-362.

[12] Sarwar B,Karypis G,Konstan J,et al. Item-based collaborative filtering recommendation algorithms [C]. New York: ACM Press,2001:285-295.

[13] 范波,程久军. 用户间多相似度协同过滤推荐算法 [J]. 计算机科学,2012,39(1):23-16.

Personalized Agricultural Products Mobile Information Service System Based on Decision Tree Classification

CHEN Yahui,YE Jihua*

(College of Computer Information Engineering,Jiangxi Normal University,Nanchang Jiangxi 330022,China)

Abstract: Aimed at mobile information services of agricultural products,combined with personalized recommendation algorithm and classification algorithm,provided a the recommend algorithm based on classification. use of decision tree classification method for classifying agricultural products,obtaining classified data,use of collaborative filtering algorithms analyze categorical data and look for similar user interest,recommend produce information to the user. Experimental results show that the traditional method and compared recommendation,the system recommended a higher degree of interest in agricultural information to mobile users.

Key words: decision tree algorithm; personalized recommendation; collaborative filtering algorithm; products-mobile information service

(责任编辑:冉小晓)