文章编号：2095-2163(2023)08-0158-04 中图分类号：TP391 文献标志码：A

基于相似度的 Apriori 混合算法研究

罗 洁[1,2]，王 力[1,3]

（1 贵州大学 大数据与信息工程学院，贵阳 550025；2 毕节工业职业技术学院，贵州 毕节 551700；
3 贵州工程应用技术学院 信息工程学院，贵州 毕节 551700）

摘 要：针对 Apriori 关联规则算法与用户兴趣度不匹配，容易造成错误商务决定的问题。提出一种新的基于相似度的 Apriori 混合算法，以提高数据分析的准确率。通过加入用户兴趣度权重，利用协同过滤算法中客观用户相似度替代主观兴趣度改进了 Apriori 算法，并进行测试。实验结果表明，改进后的算法平均置信度提高了 13%，平均支持度提高了 25%，有效提高了关联规则的准确性。

关键词：Apriori 算法；兴趣度；相似度；协同过滤

Research on Apriori hybrid algorithm based on similarity

LUO Jie[1,2], WANG Li[1,3]

(1 College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China;
2 Bijie Industry Polytechnic College, Bijie Guizhou 551700, China;
3 School of Information Engineering, Guizhou University of Engineering Science, Bijie Guizhou 551700, China)

【Abstract】 Because of the mismatch between Apriori association rules algorithm and user interest, it is easy to make wrong business decisions. A new Apriori hybrid algorithm based on similarity is proposed to improve the accuracy of data analysis. Apriori algorithm is improved and tested by adding user interest weight and replacing subjective interest with objective user similarity in collaborative filtering algorithm. Experimental results show that the improved algorithm improves the average confidence by 13% , the average support by 25% , and improves the accuracy of association rules.

【Key words】 Apriori algorithm; interest; similarity; collaborative filtering

0 引 言

大数据飞速发展的时代，各种商品琳琅满目，用户的需求与日俱增。数据挖掘就是帮助人们从大量数据中找出相关的有用信息，进而加以利用。商家就可以通过关联规则挖掘算法得到用户需求的相关规则，从而制定营销策略。目前，已有大量的学者对此进行研究。杜永兴等学者[1]通过增加判断集，减少候选项集的产生，提出了基于荒漠草原数据多样性关联规则改进的 Apriori 算法，减少时间运算。何庆等学者[2]通过改进 Apriori 算法对贫困户建档立卡数据进行挖掘。郭凯等学者[3]提出的基于 Apriori 断面越限调整策略的改进算法，大大减少了无效规则的产生。林陈[4]提出的基于兴趣度的关联规则算法有效地降低了运行时间。上述关联分析旨在改进算法，减少运行时间，未能从商家角度去分析数据结果的有效性，利用率不大。

针对基于兴趣度的 Apriori 算法存在主观因素，不利于计算关联规则的准确度。本文旨在添加相似度的属性，利用客观的相似度替代主观的兴趣度，通过用户对商品的打分，分析用户的喜好，从而得到相似度；改进 Apriori 算法得到关联规则。同时，对电影数据进行分析研究，在此基础上指导商家如何安排电影上架，验证了混合算法的有效性。

1 相关知识

1.1 Apriori 算法

Apriori 算法可通过已知的频繁项集来构成长

基金项目：贵州省教育厅创新群体重大研究资助项目（黔财教合[2016]118）；贵州省首批国家级新工科研究与实践资助项目（黔教高函[2018]209 号）。

作者简介：罗 洁（1992-），女，硕士研究生，讲师，主要研究方向：计算机视觉、数据挖掘；王 力（1971-），男，教授，主要研究方向：信息系统分析、设计与开发、数据挖掘等。

通讯作者：罗 洁 Email：1278928107@qq.com

收稿日期：2022-08-22

度更大的项集，将其称为候选频繁项集。如果 k 项集满足支持度大于最小支持阈值 min_sup，则称为频繁项集 L_K；而候选 k 项集 C_K 是指由有可能成为频繁 k 项集的项集组成的集合. 具体的实现过程可以分解为 2 部分：

(1)找出数据库对象 D 中所有大于等于用户指定的 min_sup 的频繁项集。

(2)利用频繁项集生成所需的关联规则，根据用户设定的最小可信度进行选择，产生强关联规则[5]。

1.2　兴趣度度量

兴趣度是一类新型的关联规则度量方式的挖掘模型，是以支持度-置信度模型为基础而提出的关联规则度量方法。一般的关联规则挖掘方法常常会产生大量的关联规则，但其中也包括许多用户并不感兴趣的规则。兴趣度恰好是通过用户感兴趣的专业知识和经验来筛选关联规则，挖掘最终产生的关联规则的[6]。

兴趣度可以分为以下 2 类：主观兴趣度和客观兴趣度，其中客观兴趣度等同于数据驱动，主要根据数据库中的数据以及规则或模式的形式进行定义；而主观兴趣度不仅要考虑数据，还要考虑一些人为因素的影响，属于用户驱动[6]。

1.3　相似度计算

在协同过滤算法当中，通过余弦相似度来计算用户之间的相似度。研究推得的数学公式如下：

$$sim_{cos} = \frac{z}{\sqrt{x * y}} \tag{1}$$

其中，$z = \sum_{j \in I_{m,n}} l_{m,j} l_{n,j}$，$x = \sum_{j \in I_{m,n}} l_{m,j}^2$，$y = \sum_{j \in I_{m,n}} l_{n,j}^2$；$I_{m,n}$ 表示用户 m,n 的共同评分项目集；$l_{m,j}$，$l_{n,j}$ 表示用户 m,n 分别对项目 j 的评分。余弦相似度以向量夹角余弦值来度量用户相似度，夹角越小，余弦值越大，相似度也就越高。

2　改进 Apriori 算法

2.1　算法过程

如何有效地挖掘关联规则是基于相似度的混合算法的重要任务。过程主要包括：数据抽取、挖掘建模和分析结果。基于相似度的混合算法过程如图 1 所示。

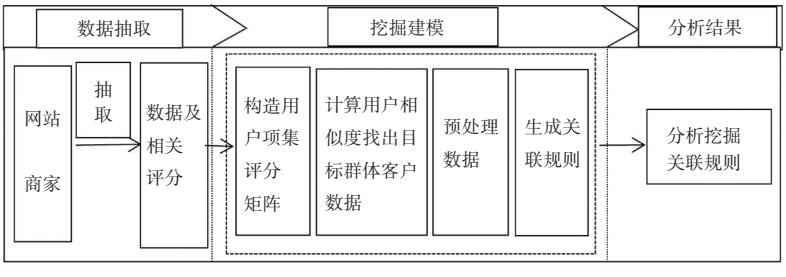

图 1　基于相似度的混合算法过程

Fig. 1　A hybrid algorithm process based on similarity

用户的打分或喜好是主观行为，但是集中相似用户的打分行为或共同喜好，就较为客观，能使目标群体更精确，得出的结论更为直接、准确，从经济的角度考虑也能在更大程度上减少商家运营成本和时间，有利于商家对一类群体进行分析及做出决策。再通过 Apriori 关联规则算法对相似用户的行为进行数据挖掘，就能得出更加有效的数据规则结果。

2.2　算法步骤

基于相似度的 Apriori 混合算法步骤具体如下。

输入　数据集、评分数据集

输出　关联规则

步骤 1　初始化数据，并对评分数据进行归一化处理。

步骤 2　找出所有数据集与用户 m 有交集用户，用式(1) 对这些用户循环计算与用户 m 的相似度。

步骤 3　根据相似度得到相似度大于等于 0.6 的列表数据。

步骤 4　对得到的列表数据进行数据预处理。

步骤 5　利用 Apriori 算法对数据推导出关联规则。

3　实验结果与分析

3.1　算法结果分析

实验环境为：Inter(R) Core(TM) i5-2410M

CPU @ 2.30 GHz; 8 GB 内存; 操作系统是 Winsdow10 64 位,利用 Jupyter Notebook 进行编程。实验数据集为 MovieLens (https://grouplens.org/datasets/movielens/)提供的 6 040 位用户对 3 925 部电影、共 1 000 209 条评论信息。从中抽取 500 位用户对 50 部电影、共 5 000 条相关评论信息进行测试。电影的评分范围为[1,5]区间所有整数值,用户对电影的喜好程度由 1~5 逐渐递增,数值越大,就表明越喜欢。抽取的每位用户对 10 部电影进行评分,实验数据集包含了用户信息、评分信息和电影信息。可以根据基于相似度的 Apriori 混合算法针对某一类人进行有效的关联规则数据分析,得出的结论更准确。实验度量指标选用了平均支持度($\overline{sup}$)和平均置信度($\overline{confd}$)。数学公式具体如下:

$$\overline{sup} = \frac{\sum_{i \in n} sup(I_i)}{N} \tag{2}$$

$$\overline{confd} = \frac{\sum_{i \in n} confd(I_i)}{N} \tag{3}$$

其中, $\sum_{i \in n} sup(I_i)$ 表示关联规则支持度的总和, N 表示规则数目。平均支持度($\overline{sup}$)越大,准确性越高; $\sum_{i \in n} confd(I_i)$ 表示关联规则置信度的总和;平均置信度($\overline{confd}$)越大,准确性越高。

(1)实验 1。为 2 种算法在最小支持度为 0.06,最小置信度为 0.75,不同用户数量条件下平均支持度、平均置信度、关联规则个数、频繁项集个数对比结果。

2 种算法在不同用户数量下的比较结果见表 1。从表 1 中看出随着用户数量的增加,Apriori 算法的平均支持度、平均置信度和关联规则个数在下降;但基于相似度的 Apriori 混合算法的平均支持度却在上升。

(2)实验 2。为 2 种算法在最小置信度为 0.75,用户数量为 500,不同最小支持度条件下平均支持度、平均置信度、关联规则个数、频繁项集个数的对比结果。

2 种算法在不同支持度下的比较结果见表 2。从表 2 中看出随着最小置信度的增加,原算法的平均支持度、平均置信度、关联规则个数、频繁项集个数都在下降,而改进算法的平均支持度却在上升。

表 1 2 种算法在不同用户数量下的比较

Tab. 1 Comparison of the two algorithms with different number of users

算法	数据集个数	平均支持度	平均置信度	关联规则个数	频繁项集个数
Apriori 算法	100	0.096	0.861	287	5
	200	0.091	0.829	123	5
	300	0.088	0.810	58	4
	400	0.079	0.795	49	4
	500	0.079	0.781	49	5
改进算法	100	0.088	0.999	48 362	10
	200	0.092	0.988	6 225	8
	300	0.107	0.910	2 575	7
	400	0.112	0.895	1 994	7
	500	0.113	0.880	1 850	7

表 2 2 种算法在不同支持度下的比较

Tab. 2 Comparison of the two algorithms under different supports

算法	最小支持度	平均支持度	平均置信度	关联规则个数	频繁项集个数
Apriori 算法	0.02	0.033	0.821	654	6
	0.04	0.061	0.792	111	5
	0.06	0.079	0.781	49	5
	0.08	0.110	0.772	12	4
	0.10	0.130	0.773	5	4
基于相似度	0.02	0.034	0.978	22 782	9
的 Apriori	0.04	0.071	0.900	4 931	8
混合算法	0.06	0.113	0.880	1 850	7
	0.08	0.145	0.863	1 068	6
	0.10	0.177	0.852	681	6

3.2 改进算法实例解析

设置最小支持度 0.6,最小置信度 0.8。关联规则结果见表 3。实例分析如下:

(1) *userid* 为 100 的用户是位男性,年龄 35 岁,职业是技术人员;从 500 位用户中查找与 *userid* 为 100 相似的用户,其中相似度为 0.6 以上的有 97 位用户;这 97 位用户中 25 位为女性,67 位为男性;年龄段为 18~56 岁,均为成年人;职业为教育行业、技术人员、大学生等有文化人员。

(2) *movieid* 为 39 的电影类型为动作/犯罪/戏剧, *movieid* 为 32 的电影类型为神秘/科幻/惊悚片, *movieid* 为 11 的电影类型为喜剧/戏剧/浪漫, *movieid* 为 21 的电影类型为喜剧/犯罪/惊悚片, *movieid* 为 34 的电影类型犯罪剧。

(下转第 164 页)