Hans Journal of Data Mining 数据挖掘, 2020, 10(2), 111-117
Published Online April 2020 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2020.102011

Hans汉斯

Research on Heterogeneous Product Recommendation Algorithm Based on Item Similarity

Jingming Chen, Fenyan Wang

School of Mathematics, China University of Geosciences, Wuhan Hubei
Email: 819277223@qq.com

Received: Mar. 9th, 2020; accepted: Apr. 1st, 2020; published: Apr. 8th, 2020

Abstract

Most of the existing recommendation algorithms are based on the recommendation of similar products, which can easily lead to "information cocoon rooms". In order to solve the limitations of similar product recommendation, the recommendation algorithm is extended to different categories of product recommendation. This paper proposes a heterogeneous product recommendation algorithm based on item similarity. Based on the application of item similarity, a cross-correlation recommendation theory is proposed to solve the problem of heterogeneous recommendation of target products and recommended product sets. Finally, this article extracts the product data from the Tianchi Taobao clothing matching data set, and applies the proposed algorithm to the programming language to analyze the data set. According to the obtained experimental results, the algorithm has a high recommendation success rate and a good recommendation effect.

Keywords

Recommendation Algorithm, Heterogeneous Products, Item Similarity

基于物品相似度的异类商品推荐算法研究

陈景明，王芬艳

中国地质大学(武汉)数理学院，湖北 武汉
Email: 819277223@qq.com

收稿日期：2020年3月9日；录用日期：2020年4月1日；发布日期：2020年4月8日

摘 要

现有的推荐算法大多是基于同类商品进行推荐，容易形成“信息茧房”。为了解决同类商品推荐的局限

文章引用: 陈景明, 王芬艳. 基于物品相似度的异类商品推荐算法研究[J]. 数据挖掘, 2020, 10(2): 111-117.
DOI: 10.12677/hjdm.2020.102011

性，将推荐算法引申到不同类别的商品推荐之中，本文提出了一种基于物品相似度的异类商品推荐算法。在应用了物品相似度的基础上，提出交叉相关推荐理论，解决了目标商品与推荐商品集异构推荐的问题。最后，本文从天池淘宝穿衣搭配数据集中提取商品数据，通过对所提出的算法设计程序语言，应用到数据集中进行分析。根据所得到的实验结果，该算法得到的推荐成功率较高，推荐效果较好。

关键词

推荐算法，异类商品，物品相似度

1. 引言

随着互联网的不断发展，人们每天产生与接收的信息量呈爆炸式增长，在无形中造成了信息过载的问题。人们迫切需要在海量信息中快速获取他们感兴趣的部分，即所谓个性化信息，于是推荐系统和推荐算法就此应运而生。推荐算法的本质是一种信息过滤技术，它根据用户对事物的偏好或者观察到的行为，通过过滤出大量的动态以及重要信息的片段，来解决信息过载的问题[1]。同时，推荐算法作为偏好学习的重要应用方向之一，其类型下的商品推荐系统已广泛应用于我们的生活之中。国内外也有很多学者对推荐算法进行了深入的研究和探讨。

江海洋(2010)提出了在评论中挖掘文字信息的新方法，将评论中用户所关注的信息挖掘出来并进行评分，从而对其他待评分对象进行预测并产生推荐[2]；李京等(2011)在研究推荐库和用户兴趣模型的基础上加入了推荐引擎，实现了基于电子商务的个性化推荐系统[3]；崔睿宇等(2019)系统地介绍了基于内容、协同过滤、基于关联规则、基于知识和基于人口统计信息等主流的推荐算法，对组合推荐算法的思路也进行了简要介绍，给出了算法评价标准并使用实验常用数据集进行了实证研究[4]；而国外也有很多学者也对此进行了深入的研究，如 Mohsen Ahmadi Fahandar 等(2017)提出基于类比推理的排序学习的方法[5]；Sandeep K. Raghuwanshi 等人(2019)针对推荐系统中的协同过滤技术进行了研究[6]；Bag, Sujoy 等人研究了使用相关的 Jaccard 相似度进行有效的推荐生成[7]。

推荐算法的出现和发展改变了信息检索和个性化推荐的模式，也有效地解决了数据过载等问题。随着推荐算法的深入研究与推广，推荐算法作为一种全新的、智能化的方法已经被广泛地使用到了生活中的各个领域之中，如商品推荐、广告推荐、浏览推荐等。然而，目前已经存在的绝大部分的推荐系统的研究主要集中在同类物品的推荐算法上[8]。为了解决不同类别的个性化推荐问题，本文提出了一种基于物品相似度的异构推荐算法，并且以异类商品的推荐为例，进行深入的研究与探讨。

2. 物品相似度

在进行推荐算法的研究时，通常会涉及相似度理论及计算，相似度可以用来衡量用户或项目之间的相似或相关程度。在推荐算法的研究中，为了将不同的物品关联起来而使用的相似度有物品相似度和用户相似度，物品相似度是用来衡量物品之间的相似度大小，而用户相似度衡量的是不同用户之间的相似度值[9]。本文需要解决的关键问题是如何将不同类别的商品联系起来，这里就采用物品相似度理论来完成异类商品的推荐。下面介绍几个相关的概念。

1) 物品相似度

物品相似度是用来衡量不同的个体之间相似性的方法，其计算方法是通过计算个体特征值之间的相似度来完成的。在这里，个体可以指商品、文本、图片等。对于不同的个体，通过特征值的形式就可以完成对个体特征的描述。物品相似度就是通过比较不同物品特征值之间的差异性或者相似性，来完成对异构项目的关联。为了度量相似性，最常用的两种度量方法分别是基于相关性和基于余弦值的相似性度量方法，即皮尔森相似度和余弦相似度。

2) 文本相似度

由于本次研究所使用数据集的特殊性，每个商品的特征值是通过如下方式给出的，首先对商品标题进行分词，分词后的词语以 ID 类脱敏处理的形式给出。因此，在本文的研究中，在衡量不同物品的相似度时，因为特征值的形式类似于文本的形式，所以使用文本相似度来衡量不同类别物品之间的相似度。

3) 文本相似度的计算步骤

简单来说，文本相似度的计算步骤表述如下：

Step1：对文本中的词语进行分词，结果以逗号隔开；
Step2：对两个不同文本中的分词结果做交集，计算共同出现的词语的次数；
Step3：计算目标文本中词语的个数；
Step4：通过文本相似度公式计算相似度值的大小。

4) 文本相似度的计算公式

文本相似度的计算公式如下：

$$s_{xy} = similarity(x, y) = \frac{Co(x, y)}{|x|} = \frac{s_i}{|x|}$$

其中，$similarity(x, y)$表示文本字段 x 和 y 的相似度函数，$Co(x, y)$表示字段 x 和 y 的分词结果共同出现的次数，$|x|$表示为字段 x 的词语的总数，s_i 为相似度参数，其中 $i = 1, 2, \ldots, n$。

3. 基于物品相似度的异类商品推荐算法

在处理异类商品推荐的时候，存在两个关键的问题，那就是相似度标准的选取和推荐成功率的计算。在这里，为适应数据集本身的参数设置，建立了以下模型来解决：

假设数据集中有商品 a，b，c，d 其中 a 和 b 是属于类别 $cat.1$ 的商品，c 和 d 则是属于类别 $cat.2$ 的商品，那么 a，b 和 c，d 就叫做异类商品。且商品 a 具有特征值 x_a，商品 b 具有特征值 x_b，商品 c 具有特征值 y_c，商品 d 具有特征值 y_d。那么，对于某一个用户而言，若该用户购买了商品 b，通过推荐算法计算出 c,d 的特征值与 b 相似，将给这名用户推荐商品 c 或者商品 d。

为了更直观的介绍本文所提出的数学模型，表 1 给出了所用到的数学符号及情况说明：

Table 1. Symbol Description
表 1. 符号说明

符号	说明
a_i, b_i	商品名称，且这些商品属于类别 1
c_i, d_i	商品名称，且这些商品属于类别 2
$x_{i_n}^{a}$	商品 a_i 的 n 个特征

Continued

$x_{i_n}^{b}$	商品 b_i 的 n 个特征
$y_{i_n}^{c}$	商品 c_i 的 n 个特征
$y_{i_n}^{d}$	商品 d_i 的 n 个特征
$u_1, u_2, ..., u_n$	用户名称
s_1, s_2	相似度参数
f_{means}	所有商品特征值数量的均值
$z(i, j)$	用户的购买行为
$Co(a_i, c_i)$	商品 a_i 和 c_i 出现相同特征值的次数
$Co(b_i, d_i)$	商品 b_i 和 d_i 出现相同特征值的次数
符号	说明
a_i, b_i	商品名称，且这些商品属于类别 1
c_i, d_i	商品名称，且这些商品属于类别 2
$x_{i_n}^{a}$	商品 a_i 的 n 个特征
$x_{i_n}^{b}$	商品 b_i 的 n 个特征
$y_{i_n}^{c}$	商品 c_i 的 n 个特征
$y_{i_n}^{d}$	商品 d_i 的 n 个特征
$u_1, u_2, ..., u_n$	用户名称
s_1, s_2	相似度参数
f_{means}	所有商品特征值数量的均值
$z(i, j)$	用户的购买行为
$Co(a_i, c_i)$	商品 a_i 和 b_i 出现相同特征值的次数
$Co(b_i, d_i)$	商品 c_i 和 d_i 出现相同特征值的次数

在本文所提出的异类商品推荐算法中，所研究的对象是商品集，即由一个商品集推荐另一个商品集的推荐算法。首先，选取一对目标商品 a_i 和 c_i。然后，需要通过相似度的计算，得到目标商品集 C 和推荐商品集 A。那么，基于物品相似度的异类商品推荐算法的研究问题可以表示如下：

输入：目标商品 a_i 和 c_i，且 a_i 和 c_i 分别属于不同的两个商品类别；

输出：相似度 s_1，s_2，推荐商品集合 A 和推荐成功率。

在基于物品相似度的异类商品推荐算法中，对于相似度和推荐成功率的计算方法如下：

1) 相似度计算

相似度的计算方法与前文介绍的文本相似度一致，由于所选数据集中不同类别的商品其特征值 f 的数目不同，因此，需要首先计算出数据集中所有商品特征值的平均值 f_{means}，再选择适当的特征 f 的个数，通过计算得到一个归一化的相似度 s_1，s_2，其公式如下：

$$S_1 = \frac{Co(c_i, b_i)}{f_{means}} = \frac{s_1}{f_{means}}, S_2 = \frac{Co(a_i, d_i)}{f_{means}} = \frac{s_2}{f_{means}}$$

2) 推荐成功率计算公式

推荐成功率表示一个用户购买了所推荐的商品的概率。其计算公式为：

$$C(u_A)=\frac{|u_A \cap u_B|}{|u_A|}$$

其中，u_A 表示购买了商品集 A 中商品的全部用户，u_B 表示所有购买了商品集 B 中商品的用户，$|u_{A\cap}u_B|$表示为既购买了商品集 A 中的商品又购买了商品集 B 中的商品的用户的数量，$|u_A|$表示所有购买了商品集 A 中的商品的用户的总数。

此时，整个测试集的推荐成功率 C_A 计算公式为：

$$C_A=\frac{1}{n}\sum_{u\in U}C\left(u_{A^*}\right)$$

其中，n 为测试集的个数。

4. 实验设计与结果

4.1. 数据集说明

为了选取符合研究条件的数据集，本文从阿里云天池大数据众智平台中，提取了天池比赛的淘宝穿衣搭配-挑战 Baseline 数据。然后根据研究需要，本文对原始数据集进行了拆分，得到两个分表格分别为商品基本信息数据和用户历史行为数据。其中，记录商品信息的表 2 仅用到了前三列的文本部分，这些是用来描述商品的特征。另外，需要特别说明的是所有 ID 类字段，包括 user_id，item_id，category_id 均已进行脱敏处理。表 2 和表 3 分别为商品信息和用户历史行为信息。

Table 2. Product information
表 2. 商品信息

列名	类型	含义	示例
item_id	bigint	商品 ID	439201
cat_id	bigint	商品所属类目 ID	16
terms	string	商品标题分词后的结果	5263, 2541, 2876263

Table 3. User history behavior
表 3. 用户历史行为

列名	类型	含义	示例
user_id	bigint	用户 ID	62378843278
item_id	bigint	商品 ID	439201
create_at	string	行为日期(购买)	20140911

实验所使用的环境为：Intel i5-8250U CPU，8G 运行内存，Win10 操作系统，MATLAB R2016a。

4.2. 数据预处理

原始数据集中存在的购买记录有 10 万余条，类型数量也较多，为了方便进行分析，首先要对数据进行预处理，这里将原数据集按照购买记录中的数目由大到小分成 4 个类别，取其中的三类数据 big_set1，big_set2 和 big_set4。它们的各项具体记录参数如表 4 所示：

Table 4. Data preprocessing results
表 4. 数据预处理结果

数据集名称	商品数/个	买家数/个	类别 ID
big_set1	59,380	101,986	368
big_set2	41,859	79,516	52
big_set4	28,388	68,541	461

此外，在原数据集中，购买记录均为淘宝的服装商品，其特征值在表 2 中 terms 这一行给出，是将淘宝网中的商品标题分词后的词语进行 ID 类脱敏后的结果。由于淘宝网中商品的标题长度不一，因此首先要计算商品的平均特征数，通过计算原数据集中所有商品特征值的平均数为：f_{means} = 14，这是实验中需要用到的重要的相似度计算参数。

4.3. 实验结果与分析

为了检验所提出的算法在实际应用中的推荐效果，本文针对数据预处理之后所提取的三个子数据集来进行实验。实验方法为选择其中一个数据集为目标集，另一个数据集为推荐集，将推荐集中的相似物品推荐给目标集中购买相似物品的用户，如果用户确实购买了该物品则表示推荐成功。为了让目标商品的购买记录尽可能的多，于是从每类别的数据集中分别提取了购买记录最多的前 5 名的商品进行分析，并将不同类别的不同商品的编号、类别等信息分别记录在各个实验结果的表格中，表 5 和表 6 分别是实验一和实验二的结果，具体结果如下：

Table 5. Results of Experiment One
表 5. 实验一结果

No.	A-ID	A-cat.ID	A-freq.	S_1	|B|	B-ID	B-cat.ID	B-freq.	S_2	|A|	C_A(%)
1	43756	368	85	4	2394	643820	52	204	4	5	80.00
2	152879	368	80	5	2241	395171	52	101	4	142	88.03
3	573509	368	78	5	1042	395171	52	101	5	26	80.77
4	327236	368	76	6	2241	395171	52	101	4	80	86.25
5	449690	368	69	4	1887	434204	52	89	4	80	85.00
平均数		\	78	5	\	\	\	\	4	\	84.01

如表 5 所示，第一列表示实验组的编号，第二列记录了从商品集 big_set1 中选取了 5 个购买记录较多的商品及这些商品的编号 A-ID，第三列则是这些商品的类别编号 A-cat.ID，第四列是该商品在数据集中出现过的次数，记为 A-freq.，第五列是相似度参数 s_1 的值，第六列为目标商品集合 B 的大小，表示为 |B|，第七列记录了商品集合 B 中商品的编号 B-ID，第八列是类别编号，记为 B-cat.ID，第九列则是 B 中的商品在 big_set2 中出现的次数 B-freq.，第十列是相似度参数 s_2 的大小，第十一列是推荐商品集合 A 的大小，也就是推荐集合中商品的个数。最后一列是计算出来的推荐成功率，用 C_A 来表示。我们主要关注的是最后一列的推荐成功率，在表 5 中，商品的平均推荐成功率为 84.01%。

Table 6. Results of Experiment Two
表 6. 实验二结果

No.	B-ID	B-cat.ID	B-freq.	S_1	|C|	C-ID	C-cat.ID	C-freq.	S_2	|B|	C_B(%)
1	573509	368	78	3	773	152924	461	216	3	0	0
2	327236	368	76	5	250	152924	461	216	4	231	91.34
3	449690	368	69	4	208	19913	461	156	3	121	76.03
4	97472	368	68	4	1663	457883	461	126	3	196	87.24
5	593852	368	59	3	773	152924	461	216	3	110	76.36
平均数		\	70	4	\	\	\	\	3	\	66.19

实验二是选取商品集 big_set1 和 big_set4 来进行推荐，在本组实验中，选取的第一个商品可能由于本身数据存在偶然性，该商品目标集中的 5 个目标商品进行推荐得到的推荐成功率为 0。计算下来，本组实验的推荐成功率的平均值为 66.19%，但在其他情况下的推荐成功率还是较高的。

5. 结论

本文针对推荐算法领域中的热点研究问题，即现阶段的推荐算法研究主要还集中在相同类别的物品推荐上这一问题，通过对异构推荐算法的研究，提出了一种基于物品相似度的异类商品推荐算法，本推荐算法的思想是通过一个目标商品集合来推荐另一个不同类的商品集合，通过实证研究表明该算法可以得到较好的推荐结果。特别地，在推荐的过程中，基于某一类的目标商品，可以同时推荐好几类其他的商品，而不是只局限于只有两个类别的推荐。综上所述，本文所提出的基于物品相似度的异类商品推荐算法是一种推荐效果较好且值得推广的推荐算法。

参考文献

[1] 杨博, 赵鹏飞. 推荐算法综述[J]. 山西大学学报(自然科学版), 2011, 34(3): 337- 350.

[2] 江海洋. 基于评论挖掘和用户偏好学习的评分预测协同过滤[J]. 计算机应用研究, 2010, 27(12): 4430-4432.

[3] 李京, 姜卫, 张跟鹏, 宋世延. 基于电子商务的个性化推荐系统研究[J]. 计算机与数字工程, 2011, 39(7): 93-97.

[4] 崔睿宇, 杨怀洲. 电子商务环境下多元个性化服务推荐研究[J]. 智能计算机与应用, 2019, 9(1): 173-177.

[5] Fahandar, M.A. and Huellermeier, E. (2017) Learning to Rank Based on Analogical Reasoning. Springer Berlin Heidelberg, 86-93.

[6] Raghuwanshi, S.K. and Pateriya, R.K. (2019) Collaborative Filtering Techniques in Recommendation Systems. In: Shukla, R., Agrawal, J., Sharma, S., Singh, T.G., Eds., *Data, Engineering and Applications*, Springer, Singapore.

[7] Bag, S., Kumar, S. and Tiwari, M. (2019) An Efficient Recommendation Generation Using Relevant Jaccard Similarity. *Information Sciences*, **483**, 53-64. https://doi.org/10.1016/j.ins.2019.01.023

[8] Luce, R.D. and Raiffa, H. (1957) Games and Decisions: Introduction and Critical Survey. Wiley, New York, NY.

[9] Brinker, K., Fuernkranz, J. and Huellermeier, E. (2006) A Unified Model for Multilabel Classification and Ranking. *European Conference on Artificial Intelligence* (*ECAI*-06), **2006**, 489-493.