

Improved Recommendation Algorithm Based on Latent Reviews Analysis

Quanmin Wang, Kaiyang Wang, Zhenguo Li, Shi Gu, Yanfeng Sun

Faculty of Information Technology, Beijing University of Technology, Beijing
Email: wangkaiyang92@126.com

Received: Mar. 12th, 2017; accepted: Mar. 27th, 2017; published: Mar. 30th, 2017

Abstract

With the coming of big data era, massive data on the Internet provide people with great convenience. But at the same time, the problem of information overloading becomes more and more obvious. At present, collaborative filtering algorithm is the mainstream recommendation algorithm to solve the problem of information overloading. However, the traditional recommendation algorithm is faced with the problem of data matrix sparsity and the problem of cold start. This paper proposes a hybrid recommendation algorithm based on features analysis and combines the idea of collaborative filtering and latent semantic analysis. Finally it recommends the optimal objects to users. The experimental results show that the improved hybrid recommendation algorithm can improve the accuracy of the recommendation results.

Keywords

Collaborative Filtering, Review Mining, Recommendation Algorithm

基于特征偏好分析的改进混合推荐算法

王全民, 王开阳, 李振国, 谷 实, 孙艳峰

北京工业大学信息学部, 北京
Email: wangkaiyang92@126.com

收稿日期: 2017年3月12日; 录用日期: 2017年3月27日; 发布日期: 2017年3月30日

摘 要

随着互联网步入大数据时代, 网络上的海量数据为人们提供极大便利的同时, 与之相伴而来的是信息过载问题。当前协同过滤算法是解决信息过载的主流推荐算法, 但传统推荐算法面临着数据矩阵稀疏性问

题和冷启动问题, 从而影响个性化推荐的准确性。本文主要研究的是基于特征偏好分析的改进混合推荐算法, 该方法将分析用户特征偏好和物品特征相结合, 再使用传统的协同过滤思想, 将最优评分对象推荐给用户。实验表明, 该算法有效地提高了推荐结果的准确性。

关键词

协同过滤, 评论挖掘, 推荐算法

Copyright © 2017 by authors and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着互联网行业迈入大数据时代, 网络为人们提供了海量的信息资源, 极大程度地满足了用户对各类信息的需求。在带来巨大便利的同时, 网络信息过载问题也变得日益突出, 用户在海量信息中找到自身所需信息的需求变得更为迫切。

在这样的背景下, 信息推荐技术作为信息过滤的有效手段, 逐渐成为计算机学术和应用领域的热门研究对象, 已经在互联网系统中被广泛地使用[1]。推荐系统针对在用户没有明确的需求场景的情况下, 通过采集用户历史信息和推荐物品的特性, 分析其中的关联特性, 推测用户对潜在物品的个性化偏好情况, 进而向用户推荐最优选择, 帮助用户选择所需要的产品, 在极大方便网站用户的同时, 也有效提高了网站的运营效果和服务质量。

在众多推荐技术中, 协同过滤算法是当前被广泛应用较为成熟的一项技术[2]。协同过滤算法认为有近似选择偏好的用户一般可能会喜欢相同的物品, 具体可以分为基于用户的、基于项目的、基于模型的三类协同过滤算法[3]。协同过滤算法的优点首先是能够过滤那些表达复杂的描述型内容, 从而解决了难以进行自动内容分析信息过滤问题, 另外还具有推荐新信息的功能。然而, 协同过滤算法在数据量不断增大的情况下也有其问题和局限性, 典型的就评分矩阵稀疏性问题和冷启动问题, 围绕这些问题, 近年来学术界进行不断深入研究, 提出了很多有效的改进方案[4]。

本文选取结合用户对产品评价内容来提高算法的准确性。不论是在资讯传媒还是电子商务领域, 用户往往会发表相关的评论, 在这些评论信息里, 同时还有隐性的肯定或否定的态度和对产品的关注度等信息, 这些评论观点对用户和网站而言都具有重要的价值, 将这些信息提取出来, 便可为用户更为精准地推荐合适的对象。

针对当前推荐算法中存在的一些问题, 本文提出通过基于用户特征分析和产品特征分析结合的改进混合推荐算法。用户初次进入系统后, 首先通过 ICT-CLAS 技术对评论文本进行分词和词性标注, 再将特征词归纳量化, 计算物品特征权重, 完成数据第一步采集处理; 下一步基于用户行为计算用户偏好特征值; 结合传统协同过滤推荐算法, 通过邻居用户聚类, 计算用户对其它项目的评价得分, 向用户推荐得分最高的项目内容。经过实验, 结果表明该方法可以提高推荐的精确度。

2. 系统概要设计

2.1. 整体框架与步骤

基于特征偏好分析的改进混合推荐算法将用户-产品评分数据作为初始数据源, 把整个推荐过程分为

以下几大步骤:

- (1) 数据初步处理: 对大量评论数据采用汉语词语分析方法对其进行分词, 同时注明词语的词性;
 - (2) 特征词归纳与量化: 筛选出频率达到特定阈值的特征词, 对其进行剪枝处理, 过滤掉不符合要求的特征词; 量化特征词维度上的评价, 并按照特定公式计算其权重;
 - (3) 构建用户偏好向量: 构建表示用户特征偏好的向量, 并按照聚类算法的思想计算用户的相似邻居集;
 - (4) 产生推荐内容: 计算用户评分相似度和特征相似度的值, 将评价分数最高的项目推进给用户。
- 简明系统框架如图 1 所示。

2.2. 相关数据

用户的集合表示为 $U\{U_1, U_2, U_3, \dots, U_M\}$; 电影产品的集合表示为 $P\{P_1, P_2, P_3, \dots, P_N\}$; 产品特征集合表示为 $F\{F_1, F_2, F_3, \dots, F_S\}$; 用户特征集表示为 $R\{R_1, R_2, R_3, \dots, R_K\}$; 用户-产品评分的矩阵用表示为矩阵 $R_{M \times N}$, 其中 R_{ij} 表示用户 U_i 对产品 P_j 的评价得分; 产品-特征关联矩阵表示为矩阵 $W_{S \times N}$ 。其中 r_{ij} 表示特征关键词 F_i 对产品 P_j 的权重比例得分。

产品-特征评分矩阵表示为矩阵 $W_{S \times N}$, 矩阵示意图如图 2 所示。

3. 新型混合推荐算法

特征偏好分析通过在文本中挖掘用户评论内容的隐含的偏好, 解决推荐算法准确性和局限性的问题。

步骤 1: 评论数据的初步分析处理; 算法首先采用汉语词法分析算法 ICT-CLAS(Institute of Computing Technology, Chinese Lexical Analysis System)对互联网上获取的电影评论文本进行中文分词[5]。ICT-CLAS 是中国科学院计算技术研究所研制的一种词语分析系统, 其主要功能包括词汇分词、词性标注等。系统

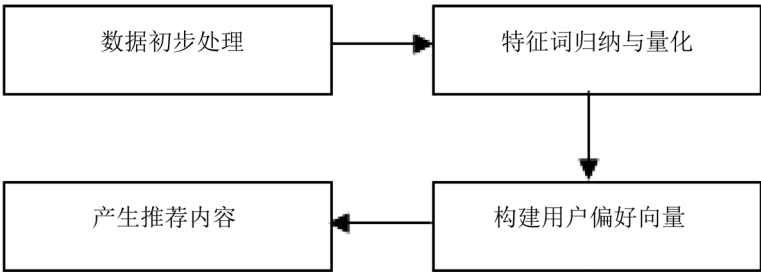


Figure 1. Framework of concise system

图 1. 简明系统框架

	Feature ₁	...	Feature _j	...	Feature _m
item ₁	r ₁₁	...	r _{1j}	...	r _{1m}
...	...	...	...	...	...
item _i	r _{i1}	...	r _{ij}	...	r _{im}
...	...	...	...	...	...
item _n	r _{n1}	...	r _{nj}	...	r _{nm}

Figure 2. Schematic diagram of product feature matrixes

图 2. 产品特征矩阵示意图

技术的基本思路是首先进行原子切分,然后在此基础上进行 N-最短路径切分,找到前 N 个最符合的切分结果,生成二元分词表,进一步进行词性标注并完成分词步骤。ICT-CLAS 系统是当前主流的汉语词法分析器。

利用汉语词语分析算法将词语注明词性,如名词、动词、形容词、副词。通过大量观察统计并设定阈值对数据进行过滤可以得出结论,特征词一般为名词和动词,情感词一般为形容词或副词,由此完成候选词中特征词与用户情感偏好词数据的采集分析。

步骤 2: 特征词的过滤与归纳; 物品特征一般指评论中被提到频率较高的描述物体性质的名词[6]。经过多次剪枝操作的尝试,得到适中的阈值标准,以此标准对候选词进行剪枝处理,对和物品性能无直接关系的候选项和小于指定阈值标准的候选项过滤,抽取和对象特征直接相关的情感词,以此提高推荐结果准确性。

具体的过滤结果如表 1 所示。

由于一种物品的特征词数量往往很多,导致建模后矩阵的维度过高计算繁琐复杂,因此可讲过滤后的特征词归纳为几个大的类型,以此简化后期计算。本文从用户评论中提取的电影特征词根据相似性可化为几类,具体划分情况如表 2 所示。

步骤 3: 构建物品特征向量; 归纳总结特征词之后,下一步是构建量化物品特征向量,将每个特征维度进行五分制评分,多个特征维度上的平均分表达用户对物品的评价倾向,最喜欢为 5 分,最不喜欢为 1 分。一个产品在特定属性上的特征值由全体用户在该特征上的共同评价决定。从而得到产品在若干个特征维度上的评价分数 $G(f)=[f_1,f_2,\cdots,f_n]$ 。部分产品特征计算结果如表 3 所示。

Table 1. Product feature words table

表 1. 过滤后部分电影特征词

过滤后部分电影特征词	
电影类型、演技、画质、摄影、视觉效果、配乐、原创音乐、音响效果、音效剪辑	导演、剧情、明星效应、美术、后期制作、剪辑、服装效果、教育意义、艺术品位

Table 2. Movie level division table

表 2. 电影特征词划分表

类型	特征词
整体	电影类型 整体评分
艺术性	导演 摄影 剧情 美术 剪辑 服装效果 艺术品味
观赏性	画质 演技 视觉效果 明星效应 后期制作
音乐	配乐 原创音乐 音响效果 音效剪辑

Table 3. Movie level scores table

表 3. 电影大类评分表

	整体	艺术性	观赏性	音乐
美人鱼	3.4	3.1	3.9	3.2
爱乐之城	4.3	4.5	4.1	4.7
乘风破浪	3.5	3.3	4.5	4.3

步骤 4: 计算特征词权重; 通过 TF-IDF 技术(Term Frequency-Inverse Document Frequency)以基于特征词频率的向量形式表示文本, 当关键词在全部文档中出现的频率较低同时在部分评论中出现的频率较高, 特征词权重较高[7]。将网页内容特征和用户资料模型相结合, 可将最终的函数定义为: 其中 D 为语料库中的文件总数, $DF(t)$ 为单词在文档中出现的频率:

$$u(c, s) = f(\text{Profile}, \text{Contents}) = tf_{i,j} * idf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \log \frac{|D|}{DF(t)} \quad (1)$$

利用公式(1)可得到 u 的数值, 并将较大的数值作为更高权重的参考特征[8]。

步骤 5: 构建用户偏好向量; 传统的推荐算法从单一维度上只考虑用户对物品的整体评分, 推荐的灵活性和准确性不高。由于不同用户对产品的关注层面各有不同, 因此本文提出一种基于用户多维度上偏好的向量推荐方法, 而且用户的爱好随着时间变化会产生变化, 通过不断地学习, 系统可以及时更新用户的特征向量。定义矩阵 R 为用户的偏好向量。 $R = [r_1, r_2, \dots, r_i]$, r_i 用户对某一大类特征的偏好程度参数。用户的特征向量计算公式为:

$$R = \frac{\sum_{i \in T} u_i * P}{\sum_{i \in T} |u_i|} \quad (2)$$

其中 P 表示产品的特征向量。用户对产品综合评分为偏好向量 R 与产品大类向量 $G(f)$ 的乘积[9]。

$$R(t) = G(f) R_i \quad (3)$$

步骤 6: 通过聚类计算用户相似度; 聚类算法通过将对象集合分为不同的子集, 相同子集中的对象具有较高的相似性, 本文通过这种思想将相似度高的用户放到一个子集中, 以此降低用户邻居集的计算量[10]。随机从用户集合中选取 k 个对象作为初始子集的中心点; 计算用户特征向量与这些子集的距离, 依据最小距离原则重新划分特征向量; 重新计算每个子集的均值作为子集的中心点; 重复前两部完成用户特征向量聚类。

用户的相似度可以分为两类——用户偏好相似度和用户评分相似度[11]。用户偏好相似度可以用夹角余弦值来进行计算, 余弦相似度计算用户偏好向量的相似度公式为:

$$\text{sim}(r_i, r_j) = \cos(r_i, r_j) = \frac{r_i \cdot r_j}{\|r_i\| \cdot \|r_j\|} \quad (4)$$

其中 r_i, r_j 表示用户的偏好向量, $\text{sim}(r_i, r_j)$ 表示两者之间的偏好相似程度, 计算结果越大表示两个用户偏好约为接近[12]。

由于偏好相似的用户具有相似的评分行为, 所以通过用户对同一物品的评分, 计算用户相似度, 选用最为常用的皮尔森相关系数公式(Pearson Correlation Coefficient), 公式如下:

$$\text{sim}(i, j) = \frac{\sum (s_{i,c} - S_i)(s_{j,c} - S_j)}{\sqrt{\sum (s_{i,c} - S_i)^2} \sqrt{\sum (s_{j,c} - S_j)^2}} \quad (5)$$

其中 $s_{i,c}$ 和 $s_{j,c}$ 表示用户 i 和用户 j 对同一评分项目 c 的评分, S_i, S_j 表示用户的平均对物品的平均评分[13]。

步骤 7: 综合评分计算和物品推荐; 利用基于用户协同过滤算法, 通过与目标用户相似的邻居集合计算目标物品对特定商品的评分, 选择相似度较大的邻居用户形成邻居集, 采用加权平均的策略, 根据

公式(7)，将邻居集中对项目的评分加权计算出推荐结果，

$$P(n,i) = S_n + \frac{\sum_{m \in \text{users}} (s_{m,i} - S_m) * \text{sim}(n,m)}{\sum_{u \in \text{users}} \text{sim}(n,m)} \quad (6)$$

其中 S_n 表示项目平均评分, $s_{m,i}$ 表示邻居用户对该项目 i 的评分, S_m 表示邻居用户所有评分均分, $\text{sim}(u, m)$ 表示目标用户 u 和邻居用户 m 的综合相似度[14]。将推荐预测评分最高的前 k 个对象推荐给用户。

4. 实验分析

实验过程

平均差值公式是目前评价一个推荐系统的性能的主流评价标准，把实际评分与预测得到的评分求绝对平均误差 MAE(Mean Absolute Error)来比较预测分值与实际分值之间的偏差程度，从而判断推荐的准确性，当平均差值越小表示推荐系统的效果越好[15]。算法预测评分集为 $\{s_1, s_2, \dots, s_n\}$ ，用户实际评分集为 $\{q_1, q_2, \dots, q_n\}$ ，平均差值的公式如下[16]，

$$\text{MAE} = \frac{\sum_{i=1}^N |s_i - q_i|}{n} \quad (7)$$

通过网络爬虫程序爬取某互联网电影资料库中用户对 1299 个产品的评价信息和评分数据，将数据随机分为两个部分，其中 80% 归为训练集，20% 归为测试集。实验结果表明，传统的基于用户或基于内容协同过滤算法绝对平均误差 MAE 值较大，推荐精度较低，相比较之下，新的推荐算法 MAE 值较小，推荐精确度有明显提高，实验数据如图 3 所示。

基于特征偏好分析的新型混合推荐算法的优势在于，首先对评论文本进行分析，提炼出其中的关键特征信息，核心思想是综合利用用户和产品两方面的特征分析方法，再通过基于用户协同过滤算法处理得到最终数据，因此新算法与传统推荐算法相比有明显优势，特别是在用户目标数较低的情况下，精确度提升效果更为明显。

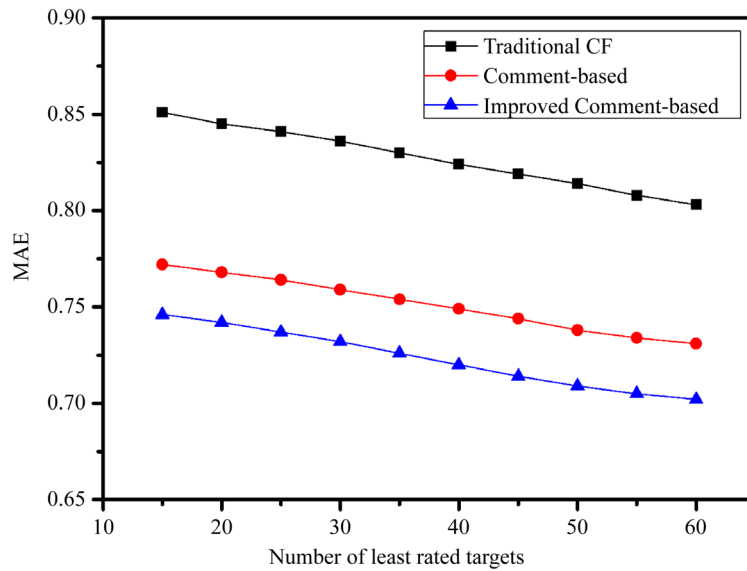


Figure 3. Comparison of algorithm recommendation accuracy

图 3. 算法推荐准确度比较

5. 结论

本文提出了一种基于特征偏好分析的改进混合推荐算法, 计算物品特征向量和用户特征偏好向量, 通过与目标用户相似的邻居用户聚类, 利用协同过滤推荐算法, 将评分最高的物品推荐给用户, 缓解了传统推荐算法中的矩阵稀疏性问题和冷启动问题。实验结果表明, 基于特征偏好分析的改进混合推荐算法有效提高了推荐结果的准确性。

参考文献 (References)

- [1] Herlocker, J.L., Konstan, J.A., Borchers, A., *et al.* (1999) An Algorithmic Framework for Performing Collaborative Filtering. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, 15-19 August 1999, 230-237. <https://doi.org/10.1145/312624.312682>
- [2] 许海玲, 吴潇, 李晓东, 等. 互联网推荐系统比较研究[J]. 软件学报, 2009, 20(2): 350-362.
- [3] 孙文爽. 岭回归参数选择的主成分方法[J]. 云南大学学报(自然科学版), 1991, 4: 1.
- [4] Bobadilla, J.S., Ortega, F., Hernando, A., *et al.* (2012) A collaborative Filtering Approach to Mitigate the New User Cold Start Problem. *Knowledge-Based Systems*, **26**, 225-238.
- [5] 项亮. 推荐系统实践[M]. 北京: 人民邮电出版社, 2012.
- [6] 杨阳, 向阳, 熊磊. 基于矩阵分解与用户近邻模型的协同过滤推荐算法[J]. 计算机应用, 2012, 32(2): 395-398.
- [7] Wu, H.C., Luk, R.W.P., Wong, K.F., *et al.* (2008) Interpreting tf-idf Term Weights as Making Relevance Decisions. *ACM Transactions on Information Systems*, **26**, 13. <https://doi.org/10.1145/1361684.1361686>
- [8] 赵丹, 蔡书. 基于网络购物评论的商品推荐模型[J]. 西南民族大学学报: 自然科学版, 2012, 38(3): 471-474.
- [9] 邓爱林, 左子叶, 朱扬勇. 基于项目聚类的协同过滤推荐算法[J]. 小型微型计算机系统, 2004, 25(9): 1665-1670.
- [10] Owen, S. and Owen, S. (2012) Mahout in Action.
- [11] 卢竹兵. 基于信任关系的协同过滤推荐策略研究[D]: [硕士学位论文]. 重庆: 西南大学, 2008.
- [12] 刘佳. 基于矩阵分解协同过滤算法的评分预测[J]. 科技创新与应用, 2016(5): 86-86.
- [13] 章诗杰, 姚俭平. 基于评论挖掘的新协同过滤推荐模型[J]. 科技创新与生产力, 2013(3): 52-54.
- [14] Chen, J. and Yin, J. (2007) A Collaborative Filtering Recommendation Algorithm Based on Influence Sets. *Journal of Software*, **18**, 1685-1694. <https://doi.org/10.1360/jos181685>
- [15] 陈志敏, 李志强. 基于用户特征和项目属性的协同过滤推荐算法[J]. 计算机应用, 2011, 31(7): 1748-1750.
- [16] 刘宇, 张广怡, 等. 协同标签系统中基于标签组合效应的推荐算法[J]. 华南理工大学学报: 自然科学版, 2013, 41(9): 65-70.

期刊投稿者将享受如下服务:

1. 投稿前咨询服务 (QQ、微信、邮箱皆可)
2. 为您匹配最合适的期刊
3. 24 小时以内解答您的所有疑问
4. 友好的在线投稿界面
5. 专业的同行评审
6. 知网检索
7. 全网覆盖推广您的研究

投稿请点击: <http://www.hanspub.org/Submission.aspx>

期刊邮箱: csa@hanspub.org