协同过滤算法的多样性研究

包增辉， 宋余庆
（江苏大学计算机科学与通信工程学院，镇江 212013）

摘要：针对现有推荐算法忽视了推荐结果总体多样性的问题，在传统的协同过滤算法基础上，提出了新的算法。分析了基于项目的协同过滤算法的原理及多样性缺陷，有针对性地对其改进。该方法降低了活跃用户、热门商品对计算商品相似度的贡献，并利用贝叶斯理论分析用户对商品特征属性的喜好度。在计算相似度时，考虑用户对商品特征的喜好度，在此基础上计算目标商品的最近邻居。实验结果表明该算法可以有效提高推荐系统的多样性。

关键词：协同过滤；多样性；贝叶斯理论；商品相似度
中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1003-8329(2013)03-0005-05

Research of Collaborative Filtering Algorithm Diversity

BAO Zeng-hui， SONG Yu-qing
(School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China)

Abstract: Aiming at the problem that the existing recommendation algorithms ignore the overall diversity, a new algorithm is presented based the traditional collaborative filtering algorithm. This paper analyzes the principles of item - based collaborative filtering algorithm and the defects of diversity, targeted for its improvement. The new algorithm reduces the contribution of active users and hot products on calculating item similarity, and analyzes the the user's preferences of the item characteristics by using Bayesian theory. It calculates item similarity degree considering the user's preferences, and then acquires nearest neighbors of the target item. Experimental results show that the new algorithm can effectively improve the diversity of the recommended system.

Key words: collaborative filtering; diversity; bayesian theory; item similarity

1 引 言

随着 Web 技术的发展，使得内容的创建和分享变得越来越容易。每天都有大量的图片、博客、视频发布到网上。信息的过载[1]使得人们找到他们需要的信息将变得越来越难。传统的搜索引擎是一个相对简单的帮助人们找到信息的工具，但搜索引擎并不能完全满足用户对信息发现的需求，而推荐系统的出现，使用户获取信息的方式从简单的目标明确的数据搜索转换到更符合人们使用习惯的上下文信息发现。

目前的推荐算法主要包括协同过滤算法[2]、基于内容的推荐算法[3]、基于网络结构的推荐算法[4,5]以及混合推荐算法[6]，评价推荐算法的指标主要有精确性、多样性、覆盖率、新颖性等。现在基

* 基金项目：江苏省普通高校研究生科研创新计划（1221170028）；江苏省高校自然科学基金资助项目(10KJB520004)。
作者简介：包增辉(1986 -)，男，硕士研究生，研究方向：推荐系统、数据挖掘。

于各种推荐方法的改进算法层出不穷，但是学者们大多只局限于改进推荐的精确性一个方面，而对多样性、新颖性、覆盖率等指标重视不足。只注重精确性的推荐方法获得的用户体验并不一定好，因为用户很可能已经知道这些热销流行的产品，得到的信息量很少，并且用户不会认同这是一种"个性化的"推荐。在改进推荐算法的多样性方面，文献[7,8]提出了一个基于协同过滤算法的改进算法，该算法在协同过滤算法得到的前 N 个商品中进行一次组合优化，找出 L 个商品（$L < N$），使得这 L 个商品两两之间平均相似度最小。这种方法虽然在应用上是有效的，但是会降低推荐的精确性。文献[9]将物理方法引入推荐，通过精巧混合能量扩散和热传导算法来提高推荐的多样性和精确性。但是这种方法的时间复杂度过高，无法进行实时推荐，在实际系统中无法应用。基于此，本文分析探讨了协同过滤算法的推荐原理，针对其多样性缺陷，提出了一种新的提高推荐多样性的算法。

2 传统协同过滤的原理及问题分析

协同过滤算法是推荐系统中最基本的算法，该算法不仅在学术界得到了深入研究，而且在工业界得到了广泛应用。其基本思想借鉴了日常在选购商品时的思路，如果自己身边的很多朋友都选购某种商品，那么自己就会很大概率的选择该商品；或者用户喜欢某类商品，当看到和这类商品相似商品并且其他用户对此类商品评价很高时，则购买的概率就会很大。协同推荐的用户模型为用户 - 商品评价矩阵，如表 1 中表示 $R_{i,j}$ 为第 i 个用户对第 j 个商品的评分。

表 1 用户评分矩阵

	商品 1	……	商品 k	……	商品 n
用户 1	$R_{1,1}$	……	$R_{1,k}$	……	$R_{1,n}$
……	……	……	……	……	……
用户 m	……	$R_{m,k}$	……	$R_{m,n}$	

协同过滤推荐一般分为两大类：基于用户的协同推荐（简称 UserCF）、基于项目的协同推荐（简称 ItemCF）。一般来说，社交网站内如 facebook 宜用 UserCF（因为用户多），而购书网站内如 Amazon 宜用 ItemCF（用户此前看过与此类似的书比某某也看过此书更令其信服，因为用户识书不识人）。下面主要分析一下 ItemCF 的计算原理及缺点。

2.1 基于项目的协同过滤算法（ItemCF）

基于项目的协同过滤算法给用户推荐那些和他们之前喜欢的物品相似的商品。其基本思路是先找到目标商品的最近邻居集合，再根据当前用户对最近邻居的评分预测当前用户对目标推荐商品的评分，然后选择预测评分最高的若干个目标商品作为推荐结果呈现给当前用户。

基于项目的协同推荐的计算主要分两步：首先是计算商品之间的相似度矩阵，查询目标商品的最近邻居，然后根据用户的历史行为，产生推荐列表。

计算商品之间的相似度常用余弦相似度，其公式为：

$$W_{i,j} = \frac{| N(i) \cap N(j) |}{\sqrt{| N(i) | | N(j) |}} \tag{1}$$

其中 $N(i)$ 是喜欢商品 i 的用户数，$| N(i) \cap N(j) |$ 是同时喜欢商品 i 和商品 j 的用户数。

在得到商品之间的相似度后，ItemCF 通过如下公式计算用户 u 对一个商品 j 的兴趣：

$$P_{uj} = \sum_{i \in N(u) \cap S(j,k)} W_{ji} R_{ui} \tag{2}$$

其中 $N(u)$ 是用户 u 喜欢的商品集合，$S(j,k)$ 是和商品 j 最相似的 k 个商品的集合，W_{ji} 是商品 j 和 i 的相似度，R_{ui} 是用户 u 对商品 i 的兴趣。

2.2 协同过滤算法的多样性问题分析

推荐系统的一大目的是给用户推荐他们不知道的、长尾中的商品。这样不仅可以提高商家的利润，而且可以提高用户使用推荐系统的体验。一般来说，用户的兴趣是多样的，比如，用户 A 喜好的电影中包含了 60% 的爱情类，20% 的冒险类和 20% 的科幻类，这种情况下推荐列表应包含 60% 的爱情类电影，20% 的冒险类和 20% 的科幻类电影。这样的推荐结果既具有一定的多样性，又考虑到了用户的主要兴趣。但是 ItemCF 算法产生的推荐列表会包含大量爱情类电影，很少甚至不出现其他两类推荐。

产生这个问题的主要原因是：

（1）每个用户对计算商品相似度的贡献是不同的。

比如，用户 A 买了当当网上 80% 的书准备用来

自己卖，用户 B 在当当网上只买了 10 本书。显然用户 B 对其买的 100 本书，计算相似度的贡献应更大一些。因为用户 A 买的书并非出自自己的兴趣，而且这些书覆盖了当当网图书的很多领域。

(2) 应减少热门商品的推荐力度。

Amazon 网站的研究人员发现 ItemCF 计算出的相似度矩阵存在一个问题，就是很多书都和《哈利波特》相关。后来他们发现，主要是因为《哈利波特》太热门了，导致它与其它商品的相似度过高，这样 ItemCF 每次都会推荐热门商品，降低了推荐的多样性和覆盖率。

(3) 只利用了用户行为信息，没有利用商品的属性信息。

ItemCF 只关注用户对商品的评价，忽视了商品种类的区别及不同类商品的属性信息。用户的兴趣是不同的，通常集中在几个领域，而 ItemCF 在计算相似度时没有考虑到用户的不同兴趣，以及不同类商品的属性信息，导致其最后推荐的多是与用户主要偏好类似的商品，推荐较为单一。

3 改进后的新算法

算法主要改进 ItemCF 在计算商品相似度这个步骤，根据以上分析的几个缺点进行优化。

3.1 降低活跃用户对商品相似度的贡献

由上面的分析，可以得出，活跃用户对商品相似度的贡献应该小于不活跃的用户，所以修正计算商品相似度的公式：

$$W_{ij} = \frac{\sum_{u \in N(i) \cap N(j)} \frac{1}{\lg(3 + K(u))}}{\sqrt{| N(i) | | N(j) |}} \quad (3)$$

其中 $K(u)$ 为用户 u 的活跃度，即用户 u 购买了多少商品，对多少商品进行了评价。

3.2 降低热门商品对商品相似度的贡献

为提高推荐结果中商品的新颖性，应降低热门商品对计算商品相似度中的权重。所以应该对公式(3)计算出来的商品相似度做些修正：

$$W_{ij} = \frac{W_{ij}}{\lg(1 + \partial p(i))} (p(i) > p(j)) \quad (4)$$

其中 $p(i)$ 为商品 i 的流行度，即被多少个用户评价过。可通过参数 ∂ 来控制最终推荐结果的新颖度。

3.3 利用贝叶斯理论提高算法的精确度

3.3.1 商品的特征属性

在实际的推荐系统中，用户对商品某类属性的偏好在短期内是稳定的，比如对服装的颜色、质地或对电影的类别、演员等。分析用户对商品特征的偏好，有利于更精确地向目标用户推荐商品。可以将商品中的属性可能具有的值，都进行编码。比如颜色属性，用 8bit 的二进位可以表示 256 种颜色。每种颜色对应一种编码。每个商品中每个特征属性的值为经过编码后的值。

3.3.2 利用贝叶斯理论改进商品相似度的精度

先把用户的评分商品分成两类：用户喜欢的商品集合和不喜欢的商品集合。然后计算用户喜欢的商品集合中某项特征值出现的概率，以及用户不喜欢的商品集合中某项特征值出现的概率，最后利用贝叶斯公式计算未评分商品具有某些特征值时用户喜欢的概率。

设事件 A 为属于用户喜欢的商品集合，事件 B 为属于用户不喜欢的商品集合，用 $m_1, m_2, \cdots, m_n$ 代表特征值，$P_u(m_i | A)$ 表示用户 u 喜欢的商品集合中出现特征值 m_i 的概率，$P_u(m_i | B)$ 表示用户 u 不喜欢的商品集合中出现特征值 m_i 的概率。

$P_u^j(A | m_i)$ 表示商品 j 具有特征值 m_i 时，用户 u 喜欢的概率。$P_u^j(A | m_1, m_2, \cdots, m_n)$ 表示商品 j 同时具有特征值 $m_1, m_2, \cdots, m_n$ 时，用户 u 喜欢的概率。根据贝叶斯公式有：

$$P_u^j(A | m_i) = \frac{P_u(m_i | A) \times P_u(A)}{P_u(m_i | A) \times P_u(A) + P_u(m_i | B) \times P_u(B)} \quad (5)$$

可将 $P_u(A)$ 和 $P_u(B)$ 都取为 0.5，那么以上公式可简化为：

$$P_u^j(A | m_i) = \frac{P_u(m_i | A)}{P_u(m_i | A) + P_u(m_i | B)} \quad (6)$$

设未评分商品 j 具有的特征值集合为 m_1，m_2，…，m_n，则由复合概率公式可得：

$$P_u^j(A \mid m_1 \cdots m_n) = \frac{P_u^j(A \mid m_1)P_u^j(A \mid m_2)\cdots P_u^j(A \mid m_n)}{P_u^j(A \mid m_1)\cdots P_u^j(A \mid m_n) + [1 - P_u^j(A \mid m_1)]\cdots[1 - P_u^j(A \mid m_n)]} \quad (7)$$

在由公式(4)计算出商品 i 和 j 的相似度后，就利用以上公式计算用户 u 喜欢商品 i 的概率，再计算商品的相似度：

$$W_{i,j} = \begin{cases} P_u^i P_u^j W_{i,j} & P_u^i > 0.5, P_u^j > 0.5 \\ (1 - P_u^i) W_{i,j} & \text{其它} \end{cases} \quad (8)$$

上述的相似度计算公式，可以使得同一个用户喜欢的两个商品的相似度具有更高的值。

4 实验评估

4.1 实验数据集及评估标准

本文采用 MovieLens 站点提供的 1M 版本数据集对文中提出的算法与传统协同过滤算法进行比较。MovieLens 是明尼苏达州立大学计算机系 GroupLens 研究小组收集的用于研究协同过滤算法的数据集，它包括 6040 个用户对 3982 部电影的 100 万个评分(评分值为 1～5)记录，每个用户至少评价了 20 部电影，并且包含了简单的用户信息和电影分类信息。

为评价算法的推荐多样性，本文使用了 Ziegler 等提出的 ILS(intra – list similarity)[8]，记 PU 为推荐算法为用户 u 产生的推荐列表，则 PU 的多样性计算如下：

$$ILS(P_u) = \frac{1}{2}\sum_{i \in P_u} \sum_{j \in P_u, i \neq j} Sin(i,j) \quad (9)$$

其中 $Sin(i,j)$ 为商品 i 和 j 的相似度，这里使用公式(1)得到的值作为其相似度。

$ILS(P_u)$ 的值越小，表明推荐列表 P_u 中的商品的种类相似性越小，推荐的多样性越好。

衡量算法的覆盖率可简单定义为推荐算法能够推荐出来的商品占总商品集合的比例。计算公式如下：

$$coverage = \frac{| \cup_{u \in U} P(u) |}{| I |} \quad (10)$$

$P(u)$ 为推荐系统给用户 u 生成的一个长度为 N 的商品列表，$| I |$ 为训练集中的商品总数。

4.2 实验结果及分析

从数据集中随机抽取 100 个用户对电影评价的记录进行分析，这 100 个用户总共对 2108 部电影进行了 12416 次评价。将每个用户的 10% 的电影评价数据作为测试集，其余作为训练算法的训练集。其中，评分在 3 分以上(含 3 分)表示用户喜欢此商品，在 3 分以下表示用户不喜欢此商品。商品的特征属性为电影的年代以及电影的 18 种类别。ItemCF 算法的邻居集合大小分别取 20～180，间隔为 20。

本文对传统协同过滤算法(ItemCF)和改进后的算法进行了比较。通过在用户数据集上进行实验，两种算法的性能比较见图 1，图 2。

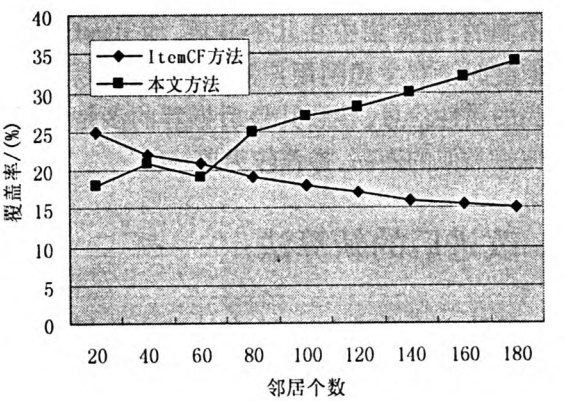

图 1 算法多样性比较

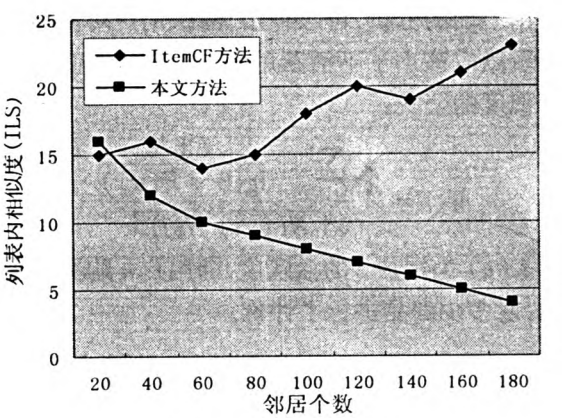

图 2 算法覆盖率比较

实验结果表明，使用改进后的算法在推荐的多样性和覆盖率方面都明显优于传统的协同过滤算法。传统的协同过滤算法随着邻居集合的增长，推荐多样性会减少，覆盖率也会缓慢递减。这是因为 ItemCF 方法在计算相似度时极易受到活跃用户、热

门商品的影响，也就是说，热门的商品会越热门，而冷门的长尾商品会越冷门。改进后的算法克服了这些缺陷，随着可参考邻居个数的增加，推荐结果会更多样化，也会有更高的覆盖率。其中的原因在于新算法在计算商品相似度时，对热门商品、活跃用户做了降权处理，也考虑到了不同用户对商品特征属性的爱好。所以使推荐结果更加精准，也更多样化。

5 结束语

用户满意是推荐系统在实际应用的终极目标，在这个总体目标下，包括了精确性、多样性、覆盖率、新颖性、实时性、商业转化等多个指标。而现在大部分的推荐技术都过多注重推荐的精确性，忽视了推荐结果的总体多样性。本文分析了传统的协同过滤的原理及多样性缺陷，对其进行了有针对性的优化，并通过实验证明了该算法对提高推荐列表的多样性有成效，可以在一定程度上提高系统的推荐质量。

参考文献

[1] BAWDEN D, HOLTHAM C, COURTNEY N. Perspectives on information overload[J]. Aslib Proceedings, 1999, 51(8): 249-255.

[2] SCHAFER J B, FRANKOWSKI D, HERLOCKER J, et al. Collaborative filtering recommender system[C]//The Adaptive Web, Lect Notes Comput Sci. Berlin, Heidelberg: Springer-Verlag, 2007, 4321: 291-324.

[3] PAZZANI M J, BILLSUS D, Content-based recommendation systems[C]//The Adaptive Web, Lect. Notes Comput Sci. Berlin, Heidelberg: Springer-Verlag, 2007, 4321: 325-341.

[4] Zhou T, Ren J, Medo M, et al. Bipartite network p rojection and personal recommendation[J]. Phys Rev E, 2007, 76: 046115.

[5] Zhou T, Jiang LL, Su RQ, et al. Effect of initial configuration on network-based recommendation[J]. Europhys Lett, 2008, 81:58004.

[6] TRAN T, COHEN R. Hybrid recommender systems for electronic commerce[C]//Proceedings of Knowledge-Based Electronic Markets, Papers from the AAAI Workshop, Technical Report WS-00-04. Menlo Park: AAAI Press, 2000: 78-83.

[7] B. Smyth, P. Mcclave, Similarity vs. diversity, in: D. W. Aha, I. Watson (Eds.). Case-Based Reasoning Research and Development[M]. Springer, 2001: 347-361.

[8] C.-N. Ziegler, S. M. Mcnee, J. A. Konstan, G. Lausen. Improving recommendation lists through topic diversification[C]. in: Proceedings of the 14th International Conference on World Wide Web, ACM Press, New York, 2005: 22-32.

[9] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling, Y.-C. Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems[C]. Proceedings of the National Academy of Sciences of the United States of America 107 (2010) 4511-4515.

(收稿日期:2013-03-11)

(上接第4页)

[5] 孙峰. 变换域自适应抗干扰方法及其FPGA实现[J]. 全球定位系统,2008,37-40.

[6] 冯冀宁,吴嗣亮,杨晓波. 一种新的干扰抑制频域自适应算法[J]. 信号处理,2010,26(12):1890-1895.

[7] 张薇薇,一种改进的快速自适应滤波算法[J]. 西安邮电学院学报,2011,16(3):6-8.

[8] 张贤达. 矩阵分析与应用[J]. 北京:清华大学出版社,2004:45-47.

[9] 刘焕淋,向劲松,代少升. 扩展频谱通信[J]. 北京:北京邮电大学出版社,2008:34-35.

(收稿日期:2013-01-23)

协同过滤算法的多样性研究

作者： 包增辉， 宋余庆， BAO Zeng-hui， SONG Yu-qing
作者单位： 江苏大学计算机科学与通信工程学院,镇江,212013
刊名： 无线通信技术
英文刊名： Wireless Communication Technology
年，卷(期)： 2013,22(3)

参考文献(9条)

1. BAWDEN D;HOLTHAM C;COURTNEY N Perspectives on information overload 1999(08)

2. SCHAFERJ B;FRANKOWSKI D;HERLOCKER J Collaborative filtering recommender system 2007

3. PAZZANI M J;BILLSUS D Content-based recommendation systems 2007

4. Zhou T;Ren J;Medo M Bipartite network p rojection and personal recommendation 2007

5. Zhou T;Jiang LL;Su RQ Effect of initial configuration on network-based recommendation 2008

6. TRAN T;COHEN R Hybrid recommender systems for electronic commerce 2000

7. B.Smyth;P.Mcclave Similarity vs.diversity 2001

8. C.-N.Ziegler;S.M.Mcnee;J.A.Konstan;G.Lausen Improving recommendation lists through topic diversification 2005

9. T.Zhou;Z.Kuscsik;J.-G.Liu;M.Medo,J.R.Wakelíng,Y.-C.Zhang Solving the apparent diversity-accuracy dilemma of recommender systems 2010