基于资源特征的协同过滤算法的研究

王 伟,徐德智,廖晖寰
(中南大学 信息科学与工程学院,湖南 长沙 410083)

摘 要:以往的协同过滤推荐算法具有数据稀疏性问题,而对于新资源还具有“冷启动”问题。为此提出了一种基于资源特征的协同过滤推荐方法。通过收集和分析用户的行为,将用户对于资源的喜好转化为用户对于关键词的兴趣权重,将用户兴趣的改变表示为用户兴趣关键词权重的改变,以此来建立和更新用户兴趣模型。最后,通过发现用户兴趣模型与资源模型之间的联系从而达到资源推荐的目的。实验表明,该算法不仅可以跟踪用户的兴趣变迁,而且没有数据稀疏性问题和新资源的“冷启动”问题。

关键词:智能推荐;个性化推荐;推荐引擎
中图分类号:TP311 **文献标识码**:A **文章编号**:1674-7720(2012)17-0004-03

Research on collaborative filtering algorithm based on resource characteristics

Wang Wei, Xu Dezhi, Liao Huihuan
(College of Information Science and Engineering, Central South University, Changsha 410083, China)

Abstract: Previous collaborative filtering recommendation algorithms have data sparsity and cold-start problems for new resources. To solve these problems, this paper put forward a method of collaborative filtering recommendation which based on resource characteristic. In order to establish and update user interest model, it collected and analyzed user behavior, and then it turned user's preference for resources into user's interestingness to keywords and turned the changes of user interest into the changes of the weights of user's interest keywords. At last, it recommended resources to user through finding the relationships among user interest models and resource models. The test shows that this algorithm can not only find the changes of user interest, but also solve data sparsity and cold-start problems of new resources.

Key words: intelligent recommendation; personalized recommendation; recommendation engine

随着网络的迅速发展,资源数量也成倍地增长。所面临的问题已经不是如何找到资源,而是怎样从资源海洋中找到自己所需要的资源。用户获取所需资源最常用的手段就是搜索关键词和浏览推荐资源。以往简单的搜索和推荐资源并没有考虑用户的个性化需求(即没有针对性),找到的资源可能与用户需要的资源差距很大。此外,有时候用户也无法准确地把自己的需求形象地表示出来。

所谓推荐引擎,就是不需要用户额外的劳动,就可以根据用户的个性化特征推测用户可能感兴趣的资源,然后再将其推荐给用户。个性化推荐在某些领域已经取得了成功,最有名的有亚马逊推荐系统、Pandora 音乐推荐系统等。目前,个性化服务的研究已经越来越受重视,尤其是在电子商务领域和搜索引擎领域。

1 相关研究

目前,针对推荐引擎的理论已经有很多研究,推荐主要可以分为基于内容的推荐、协同过滤推荐和混合推荐。协同过滤推荐又可分为基于用户的推荐、基于项目的推荐和基于模型的推荐。参考文献[1]中论述了推荐引擎的工作原理和其中涉及的各种推荐机制。参考文献[2]和[3]中论述了在协同推荐算法中加入了用户背景信息,将用户或者资源进行分类以提高推荐的准确度。参考文献[4]在协同推荐算法中加入时间因素以跟踪用户的短期兴趣和长期兴趣。以往的协同推荐算法都是根据

用户以往对于资源的兴趣评分来推测该用户对其他未评分的物品的兴趣评分，它只考虑用户对物品的态度，而忽略了物品本身的属性和特征，因此对于新物品的推荐有“冷启动”问题。此外，它还具有数据稀疏性问题。

针对以往协同过滤推荐算法的不足，本文提出了基于资源特征的协同过滤推荐算法。通过记录和分析用户在网站上的动态行为，将用户对于资源的喜好转化为用户对于关键词的兴趣权重，将用户兴趣的变化转化为用户兴趣关键词权重的变化，以此建立用户兴趣模型。最后，通过建立用户兴趣模型与资源模型间的关联达到资源推荐的目的。它不仅没有“冷启动”问题和数据稀疏性问题，而且能够跟踪用户的长期兴趣和短期兴趣。

2 相似度策略

常用的相似度计算方法主要有欧氏距离、余弦相似性、相关相似性和修正的余弦相似性。本文采用余弦相似性[5]方法计算两个空间向量的相似度。

设用户 U_1 的关键词集合为 A，U_2 的关键词集合为 B。如果 U_2 为用户，则取集合 A 和 B 的并集作为标准关键词集合 S，即 $S=A\cup B$；如果 U_2 为资源，则取集合 B 作为标准关键词集合 S，即 $S=B$。

设 U_1 对应于 S 的权重向量为 x，U_2 对应于 S 的权重向量为 y，则 x、y 为 n 维项空间上的向量。x 与 y 之间的相似性可以通过向量间的余弦夹角度量。因此 U_1 和 U_2 的相似性 $\mathrm{Sim}(U_1,U_2)$ 为：

$$\mathrm{Sim}(U_1,U_2)=\cos(x,y)=\frac{x\times y}{\|x\|\times\|y\|}$$

式中，分子为两个向量的内积，分母为两个向量模的乘积。

3 基于资源特征的协同推荐算法

本文提出的基于资源特征的协同推荐算法以用户对于所有兴趣关键词的权重向量来描述用户，以最喜欢目标资源的多个用户的兴趣权重向量来描述目标资源，通过计算目标资源向量与其他资源向量之间的相似度来查找与该资源最相似的资源，从而达到推荐的目的。整个推荐流程如图 1 所示。

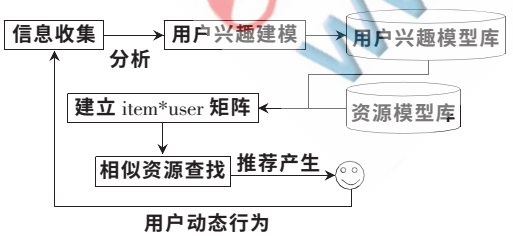

图 1 基于资源特征的协同推荐模型

3.1 信息收集

本文的信息收集不同于以往的协同推荐算法，它通过收集用户在网站上的动态行为来作为用户的兴趣源。以基础教育资源网为例，能够表达用户爱好的操作行为主要有浏览、播放、下载、预览、推荐、收藏、删除收藏、分享、搜索、评分、评论、购买等。不同的行为所表达的用户对于资源的爱好程度不一定相同（例如浏览和收藏表达的用户爱好程度不一致）。因此，当用户执行该类操作时，需要记录用户操作的类型和访问时间作为用户兴趣的依据。

3.2 用户兴趣建模

考虑到网站的性能需求，用户兴趣模型的更新是周期性的，即离线进行。用户兴趣模型的建立和更新分为以下几个步骤：

（1）将用户行为记录转化为用户关键词兴趣权重，并把对应关键词的最后访问时间设定为该行为的发生时间，然后删除该行为记录。在将用户的行为转化为用户兴趣关键词权重时，根据行为的不同对应关键词的权重增量也不同，例如浏览时与资源相关的关键词的兴趣权重分别增加 a，而收藏时与资源相关的关键词的兴趣权重分别增加 $2a$，删除收藏则对应关键词权重增量为 $-2a$。关键词兴趣权重值最大不应超过 $W\max$（最大权重值 $W\max$ 为常数），且不能小于 0（小于 0 则删除该记录）。

（2）根据时间窗（为一常数）更新所有兴趣关键词权重。用户的兴趣可能会随着时间的变化而变化，对于那些用户不再感兴趣的关键词，其兴趣权重应下降。因此，如果当前时间与某关键词的访问时间之差大于时间窗 t 时，则对应关键词的权重 W 会减少 b（b 为常量），如果 $W\leqslant 0$，则删除该关键词记录。

（3）以用户为单位采用极差变换法标准化用户兴趣关键词权重。因为通过以上步骤获得的用户兴趣模型是不标准的，需要进行标准化处理之后才能正确分析出用户的兴趣。

3.3 推荐的产生

推荐结果的产生可以分为以下几个步骤（相似度计算采用本文第 2 节介绍的余弦相似度计算方法）：

（1）建立矩阵 $A=(a_{ij})_{m\times n}$，其中 m 为资源数量，n 为最喜欢目标资源的前 n 个用户。矩阵的第 i 行记为 A^i。

（2）计算目标资源 R 与所有用户兴趣模型的相似度，相似度最高的前 n 个用户（也可以取相似度大于某个临界值的所有用户）即为最喜欢该资源的前 n 个用户。设最喜欢目标资源 R 的用户集合 $V=\{v_1,v_2,\cdots,v_n\}$，目标资源 R 与用户 $V[i]$ 的相似度为 $\mathrm{Sim}(V[i],R)$，其中 $V[i]\in V$。设 $A^0=\mathrm{Sim}(V[i],R)$，其中 $i=0,1,\cdots,N-1$。

（3）分别计算用户 $V[i]$ 的兴趣模型与其他所有资源模型的相似度。设用户 $V[i]$ 对资源 j 的相似度为 $\mathrm{Sim}(V[i],j)$，则 $a_{ij}=\mathrm{Sim}(V[i],j)$，其中 $V[i]\in V$；$i=0,1,\cdots,n-1$；$j=1,\cdots,m-1$。

（4）计算目标资源与其他资源之间的相似度。矩阵的每一个行向量都表示一个资源，其中 A^0 为目标资源的向量。通过计算矩阵 A^0 与 $(A^i)^{\mathrm{T}}(i=1,2,\cdots,m-1)$ 的余弦相似度，选取相似度最高的前 k 个资源即为与目标资

源最相似的资源，也就是推荐的资源列表。

4 实验结果和分析

4.1 实验数据集

本文基于北京国之源公司提供的基础教育资源测试数据集对上述算法的有效性进行了测试，并与传统的协同过滤推荐算法进行了比较。此数据集包含各类数据共 9 万多条，数据集采用高中一年级的语文资源数据约 3 000 条，测试用户数量为 100，每个用户至少访问过 30 个资源。

4.2 度量标准

推荐质量的评价标准采用平均绝对误差 MAE（即通过计算预测的用户评分与实际的用户评分之间的误差）来度量，MAE 值越小，推荐质量越高。

用户 u 对于目标资源 R 的真实评分 $P_{u,R}$ 可表示为：

$$P_{u,R}=5\times \mathrm{Sim}(u,R) \tag{1}$$

式中，$\mathrm{Sim}(u,R)$ 为用户 u 与目标资源 R 的余弦相似度。

设目标资源 R 的最近邻集合为 $N_p=\{r_1,r_2,\cdots,r_n\}$，资源 R 与资源 r_i 的相似度为 sim (R,r_i)（其相似度计算按第 3.3 节的步骤进行），其中 $r_i\in N_p$。则用户 u 对于资源 R 的预测评分 $Q_{u,R}$ 可表示为[6]：

$$Q_{u,R}=5\times\frac{\sum_{i=1}^{n}(\mathrm{Sim}(u,r_i)\times\mathrm{Sim}(R,r_i))}{\sum_{i=1}^{n}|\mathrm{Sim}(R,r_i)|} \tag{2}$$

式中，$\mathrm{Sim}(u,r_i)$ 为用户 u 与资源 r_i 的余弦相似度。

设预测的用户评分集合为 $\{p_1,p_2,\cdots,p_n\}$，对应的用户实际评分集合为 $\{q_1,q_2,\cdots,q_n\}$，则平均绝对误差 MAE 可表示为：

$$\mathrm{MAE}=\frac{\sum_{i=1}^{n}|p_i-q_i|}{n} \tag{3}$$

4.3 实验结果

通过对本文所提出的基于资源特征的协同过滤算法进行测试和与传统的协同过滤推荐算法进行比较可知，本文算法 MAE 值比传统算法低。实验结果如图 2 所示。

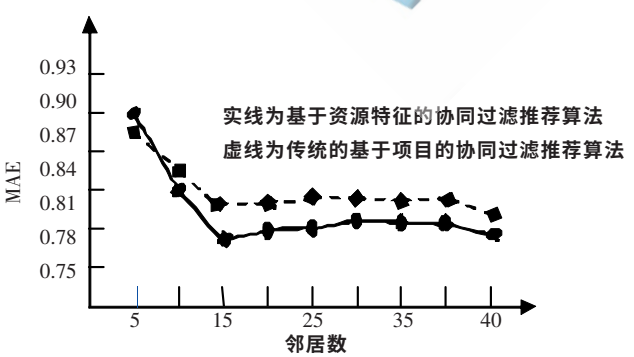

图 2 基于项目特征的协同过滤推荐算法

从图中可以看出，本文的基于资源特征的协同过滤推荐的准确性要比传统的基于项目的协同过滤推荐算法高；邻居数太少，会使推荐的准确率降低，而邻居数太多，则对推荐的准确性影响不大。

4.4 实验结果分析与比较

本文所提出的基于资源特征的协同过滤推荐算法与传统的基于项目的协同过滤推荐算法的主要不同点在于用户兴趣的表现方式不同。传统的基于项目的协同过滤推荐算法是以资源整体为单位来表示用户的兴趣，而基于项目关键词的协同过滤推荐算法是以资源特征为单位来表示用户的兴趣。

与传统的基于项目的协同过滤推荐算法相比，本文所提出的基于资源特征的协同过滤推荐算法可以跟踪用户的短期兴趣和长期兴趣，不存在数据稀疏性问题和新资源的“冷启动”问题，所需的显示用户反馈比较少，但是计算的复杂度比传统算法高。

本文根据以往协同推荐算法的不足，提出了一种基于资源特征的协同过滤推荐算法。通过在基础教育资源网上的实验结果表明，该算法解决了数据稀疏性问题和新资源的“冷启动”问题。同时，它还能够跟踪用户的兴趣变迁，而推荐质量也有所提高。下一步的工作是研究根据用户的背景和用户的关键词兴趣模型对用户进行聚类，以减少相似资源的计算开销并提高推荐的准确性。

参考文献

[1] 赵晨琳，马春娥.探索推荐引擎内部的秘密，第 1 部分：推荐引擎初探 [EB/OL].(2011-03-16)[2012-03-02]. http://www.ibm.com/developerworks/cn/web/1103_zhaoct_recommstudy1/.

[2] 吴一帆，王浩然.结合用户背景信息的协同过滤推荐算法[J].计算机应用，2008，28(11)：2972-2974.

[3] 刘旭东，葛俊杰，陈德人.一种基于聚类和协同过滤的组合推荐算法[J].计算机工程与科学，2010，32(12)：125-127.

[4] 战守义，井新.加入时间因素的个性化信息过滤技术[J].北京理工大学学报，2005，25(9)：782-785.

[5] 曾子明，于小鹏. 电子商务推荐系统与智能谈判技术[M].武汉：武汉大学出版社，2008：30-118.

[6] SARWAR B, KARYPIS G, KONSTON J, et al. Item-based collaborative filtering recommendation algorithms[C]. In: Proceedings of the 10th international conference on World Wide Web, 2001:285-295.

（收稿日期：2012-03-02）

作者简介：

王伟，男，1987 年生，硕士，主要研究方向：个性化服务。

徐德智，男，1963 年生，教授，主要研究方向：Web 计算，本体映射。

廖晖寰，男，1988 年生，硕士，主要研究方向：语义网。