基于标签、得分和偏好时效性的项目推荐方法①

张秀杰，朱克珊，李 钢

(天津大学 管理与经济学部，天津 300072)

摘 要：网络信息的爆炸式增长使得推荐系统成为一项研究的热点。现存的推荐系统在实际运营中存在各自的缺陷。在 web2.0 环境下，标签、项目得分以及用户标注项目的时间均包含暗示用户偏好的重要信息，这些信息对提高推荐系统准确度是十分重要的。在借鉴协同过滤思想的基础上，提出综合考虑标签、项目得分和用户偏好时效性的项目推荐模型，并对此模型的体系结构及应用前景进行了分析。

关键词：标签；得分；偏好时效性；协同过滤；项目推荐

Item Recommendation Method Based on Tag, Rating and Preference Timeliness

ZHANG Xiu-Jie, ZHU Ke-Shan, LI Gang

(Department of Management and Economics, Tianjin University, Tianjin 300072, China)

Abstract: The explosive growth of network information system makes the recommendation become a research hotspot. Existing recommendation systems used in the actual operation have respective defects. In web2.0 environment, tags, the item rating and the time of the user tagging the item contain important information suggesting the user's preference. This information is useful to improve the accuracy of the recommendation system. Drawing on collaborative filtering method, we suggest an item recommendation model which is considering tags, the item rating and timeliness of user's preference all together and discuss the architecture and prospect of this method.

Key words: tag; rating; preference timeliness; collaborative filtering; item recommendation

1 引言

目前，网络信息呈现了爆炸性的增长趋势，用户在众多的信息资源中，很难找到自己喜欢的项目。传统的搜索算法只能呈现给所有用户同一的排序结果，无法针对不同用户的兴趣爱好提供个性化服务。个性化推荐是解决这一问题的最有效工具之一。Web2.0 的发展使得用户与互联网的交互成为可能，在这一技术的推动下，个性化推荐研究得到了迅猛发展。根据推荐算法的不同，推荐系统可以分为如下几类：协同过滤推荐（collaborative filtering recommendation, CFR）系统；基于内容（content-based）的推荐系统；混合（hybrid）推荐系统以及基于用户-产品二部图网络结构（network-based）的推荐系统[1]。协同过滤系统是目前应用最广泛的个性化推荐系统。CFR 的基本思想是：首先为目标用户寻找兴趣相似的邻居用户，然后把邻居用户感兴趣的项目推荐给目标用户。CFR 的算法又可以分为两类：基于记忆（memory-based）的和基于模型（model-based）的算法[1]。CFR 最大的优点是能够处理非结构化的复杂对象。但是在实际应用中，CFR 仍存在一些问题，比如数据稀疏性等。这影响了推荐系统的准确度。

在 web2.0 中，标签（tags）是一项重要的信息资源，它既表达了资源项目的主要特征，又涵盖了用户与资源之间、用户与用户之间的关系，兼具内容与关联的特征。由于标签的这些特点，使其常被应用于个性化推荐领域。将标签作为推荐技术的数据来源，便有可能开发出同时具备内容过滤和协同过滤优越性的推荐技术。但是，因为标注项目的标签都是用户随意定义的，所以就不可避免地包含许多噪音。这些噪音

① 基金项目:科技部纵向课题资助(2009GJA10046)
收稿时间:2011-07-06;收到修改稿时间:2011-08-17

给提高项目推荐的准确度带来了困难。为排除这些噪音标签，有必要综合考虑多方面的因素来发现不同用户之间确实存在的偏好相似性。若以用户对项目的得分因素为桥梁，将用户标签所对应的不同项目引入到模型中，在此基础上进行用户相似性的判断，则可以避免此类错误的发生。同时，用户对项目的评分非常直观地反映了用户对某一项目的喜爱程度，从这方面来看，将标签与用户对项目的评分集成也有助于提高推荐系统的准确度[2]。另外，用户兴趣对时间是敏感的，从心理学的角度来说：①类似于记忆，人们的兴趣是随着时间的推移而逐渐下降的；②忘记的速度逐渐变慢，积累的兴趣会越来越稳定[3]。在进行个性化的项目推荐时，有必要将时间因素加以考虑。综上，标签、用户对项目的评分以及用户标注项目的时间包含反映用户兴趣及其变化趋势的重要信息，综合考虑标签、项目得分和用户偏好时效性对提高推荐系统的准确度是有帮助的。

本文借鉴协同过滤推荐算法思想，提出基于标签、得分和偏好时效性（tag, rating and preference timeliness, TRPT）的项目推荐方法，以此来提高个性化推荐系统的准确度，并对此模型的体系结构及应用前景进行分析。

2 相关文献综述

近十年来，推荐系统成为比较活跃的研究领域[4]。目前，已有众多基于协同过滤的项目推荐系统的研究[2,5,6]。

Liang Huizhi 等人[6]提出了利用用户、项目和标签之间的多重关系来发现每个用户每个标签语义含义、从而得到每个用户最可能感兴趣的项目的方法，来提高推荐系统的准确度。此方法主要解决用于协同推荐的标签中常含有许多噪音因素（比如同义标签、语义标签、私人标签）这一问题。并指出在今后的研究中，可以集成更多的用户信息，比如回顾、日志以及明确的得分，来进一步提高项目推荐系统的准确度。Nan Zheng 等人[7]研究了在社会化标注环境下，使用标签和时间信息来预测用户偏好的重要性，建立了基于这些信息的项目推荐模型，并在实验中证实了将标签和时间信息集成到协同过滤推荐系统中可以提高系统的准确度。但是，此方法忽视了用户对项目评分这一反应用户兴趣的重要因素。Heung-Nam Kim 等人[8]提出通过建立个人用户模型的方法来提高个性化推荐系统的准确度。此方法将标签和得分结合，得到用户兴趣主题，将协作方法集成到基于内容过滤方法中，从而提出了能减轻“冷启动”和过于专门化问题的推荐系统。但是，此方法中的用户兴趣主题没有考虑到兴趣的时间属性。

以上研究中，考虑了标签、项目得分和用户偏好时效性中的一个或两个因素，这样不可避免地遗漏了显示用户当前兴趣的部分重要因素，从而影响推荐系统的准确度。因此，本文借鉴协同推荐的思想，提出基于 TRPT 的项目推荐方法。

3 基于TRPT的项目推荐方法

基于 TRPT 的项目推荐方法，综合和反映用户偏好的多重信息，其项目推荐过程模型如图 1 所示。

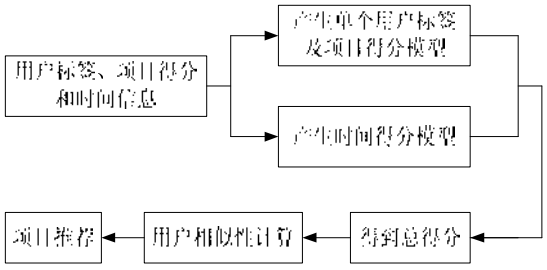

图 1 项目推荐过程模型

3.1 建立单个用户标签及项目得分模型

综合考虑用户标注的标签和用户对项目的评分，建立单个用户标签及项目得分模型。用户、标签、项目和单个用户对单个项目的评分之间的关系如图 2 所示。

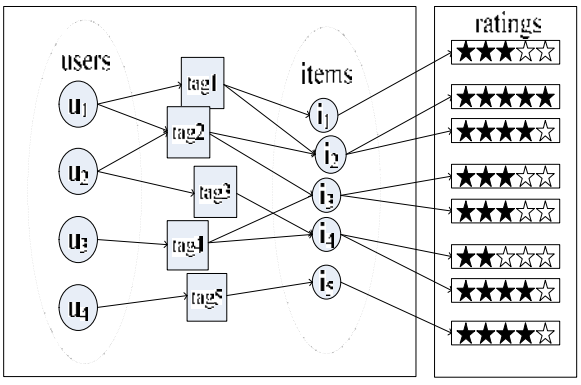

图 2 用户、标签、项目以及项目得分关系

设所有用户的集合为 $U=\{u_1,u_2,...,u_U\}$，所有项目的集合为 $I=\{i_1,i_2,...i_I\}$，所有用于标注项目的标签的集合为 $T=\{t_1,t_2,...,t_T\}$。对于项目的得分，采用了许多研究[8,9]均采用的 5 分制评分法。用户 u 对项目的评价分数为 $r_{u,i}$（$r_{u,i}$ 为 1 到 5 的整数）。

通过综合考虑项目得分和用户对项目的标注，可以得到用户 u 用于标注项目 i 的标签 t 的得分，将其表示为 $w_{u,i}(t)$，其计算公式为：

$$w_{u,i}(t)=\frac{r_{u,i}}{\sqrt{\sum_{j=1}^{I} r_{u,j}^2}} \tag{1}$$

综合考虑用户标签和用户对项目的得分因素之后得到的结果 $w_{u,i}(t)$ 是介于 0 和 1 之间的。

3.2 建立时间得分模型

综合考虑标注项目的时间和用户标注行为的生命周期两方面来对时间因素进行得分量化：

$$w_{time}(u,i)=\exp\{-\ln 2\times time(u,i)/hl_u\} \tag{2}$$

其中，$w_{time}(u,i)$ 为反应用户 u 对项目 i 的喜爱程度的时间得分。$time(u,i)$ 是一个非负整数，当项目 i 是被用户 u 在最近的一个标注日期中进行的标注时，取值为 0，当为用户 u 在倒数第二个标注日期中进行的标注时，取值为 1，依次类推；对同一个标注日期进行的标注，$time(u,i)$ 的取值是相同的。对于一个给定的用户，当进行标注的时间与最近的标注日期越接近时，$time(u,i)$ 的值越小。hl_u 代表用户 u 的半衰期，这个是与用户 u 标注行为的生命周期相对应的，用户的 hl_u 越大，代表其标注行为的生命周期越长，兴趣下降速度越缓慢，反之类似。当 $time(u,i)=hl_u$ 时，$w_{time}(u,i)$ 降低到 1/2。因此，新的标签将被赋予一个较高的时间得分值，旧的标签将被赋予一个较低的时间得分值。

3.3 综合标签得分和时间得分，得到总得分

定义一个用户-项目矩阵，用来表示用户对 u 项目 i 的偏好总得分，此矩阵表示为：

$$M(U\times I)=\begin{bmatrix} M_{1,1} & M_{1,2} & ... & M_{1,I} \\ M_{2,1} & M_{2,2} & ... & M_{2,I} \\ ... & ... & ... & ....\circ \\ M_{U,1} & M_{U,2} & ... & M_{U,I} \end{bmatrix}$$

其中 U 为用户个数，I 为资源个数，$M_{u,i}$ 为综合考虑了标签及项目得分和时间得分的用户 u 对项目 i 的偏好总得分，其表达式为：

$$M_{u,i}=\lambda w_{u,i}(t)+(1-\lambda)w_{time}(u,i) \tag{3}$$

其中的参数 λ 是介于 0 和 1 之间的，标签和项目得分更多反映用户的偏好，而时间更多反映用户偏好的变化，所以 λ 的具体取值可根据这两个因素的重要程度进行调整。当认为提高最佳平均点击率（best average Hit-rate）更重要时，λ 取较大的值；当认为提高最佳平均点击排名（best average Hit-rank）更重要时，λ 取较小的值。一般情况下，可取 λ =0.5，即认为二者同等重要。

3.4 用户相似性的计算

用户相似性的计算有很多方法[1]，最常用的有 Pearson 相关性和夹角余弦，两种方法均定义用户 u 和用户 v 共同打分的项目集合为 $I_{u,v}=I_u\cap I_v$。夹角余弦方法用两个向量之间的相似性来表示用户之间的相似性，更适合本文的研究方法。因此，本文采用夹角余弦来计算用户之间的相似性。

用户 u 和用户 v 都用相同维数的向量表示，两个向量之间的相似性可以通过计算它们之间的余弦值得到：

$$sim(u,v)=\frac{\vec{u}\times\vec{v}}{||\vec{u}||\times||\vec{v}||}=\frac{\sum_{x\in I_{u,v}} M_{u,x}\times M_{v,x}}{\sqrt{\sum_{x\in I_{u,v}} M_{u,x}^2}\times\sqrt{\sum_{x\in I_{u,v}} M_{v,x}^2}} \tag{4}$$

用户 u 与用户 v 相似性得分 $sim(u,v)$ 值越大，表示他们的兴趣越接近。取与用户 u 兴趣最接近的 M 个用户作为他的邻居用户集，表示为 $Neighbor(u)$。

3.5 项目推荐

对于一个给定的目标用户 u，通过上述步骤已经得到与其偏好相似的 M 个用户及其偏好信息。那些被邻居用户所喜欢的（偏好总得分较高的）、而又没有被目标用户所了解的项目，可以作为候选项目推荐给目标用户。推荐过程中，综合考虑邻居用户与目标用户的相似度以及邻居用户对项目的偏好总得分这两个因素。

基于以上思想，综合考虑与目标用户 u 相似的多个用户的偏好信息，进而对目标用户可能偏爱的项目进行预测，其得分表示为：

$$score(u,i_x)=\frac{\sum_{v\in Neighbor(u)} M_{v,i_x}\times sim(u,v)}{|\sum_{v\in Neighbor(u)} sim(u,v)|} \tag{5}$$

其中的 $Neighbor(u)$ 表示目标用户 u 的邻居用户集，i_x 为已被用户 v 评分而未被用户 u 评分的项目。取得分

最高的前 N 个项目作为目标用户 *u* 的推荐项目。

4 基于TRPT的项目推荐系统结构及应用前景

4.1 系统结构

根据基于 TRPT 的项目推荐方法的推荐过程，构建项目推荐系统结构。推荐系统的基本框架由用户接口、推荐服务和数据库三大部分组成。如图 3 所示。

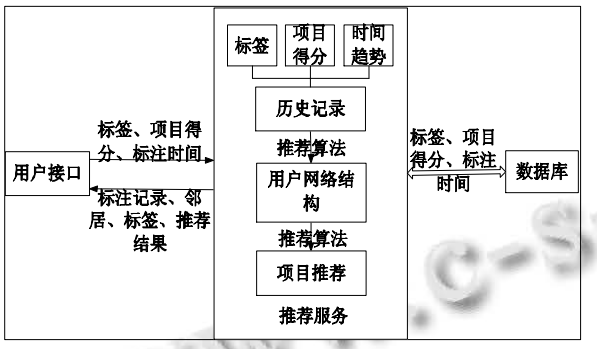

图 3 推荐系统结构

4.1.1 用户接口

用户通过用户接口（user interface）与系统进行交互，接口界面显示用户标注的历史记录、用户邻居、用户标签以及项目推荐结果。此模块的信息由推荐服务模块提供。

4.1.2 推荐服务

历史记录以图表的形式记录用户所用的标签、项目的得分以及标签的使用频率随时间的变化趋势。

用户网络结构中记录当前用户的邻居用户信息。通过对邻居用户偏好信息进行提取，可以得到当前用户潜在偏爱的项目。

项目推荐是本系统的核心，记录当前用户潜在的项目偏好信息。

推荐服务模块通过对历史记录信息进行处理，得到当前用户的邻居用户集，之后根据 TRPT 推荐算法中的规则，对邻居用户的偏好信息进行提取，进而得到当前用户可能感兴趣的项目集合。

4.1.3 数据库

数据库中存储用于项目推荐的用户的信息、标签、标注的项目信息、时间信息以及得分情况。推荐服务模块通过分析调用此模块中的数据来实现系统的推荐功能。

推荐服务模块一端连接用户接口，另一端连接数据库。当用户产生一个页面请求时，推荐服务模块会按约定的规则对数据库模块中的信息进行调用和处理，并将处理的结果返还给用户，从而实现项目的推荐。此系统可以在 Windows 操作系统中用 JSP 和数据库等相关技术实现。

4.2 应用前景

基于 TRPT 的项目推荐方法从标签、项目得分和用户偏好时效多维度进行综合，以用户对项目的得分因素为桥梁，将用户标签所对应的不同项目引入到模型中，解决了标签中常含有许多噪音因素这一问题。考虑了之前多数方法均没有考虑的偏好时效性问题，将最近标注的项目赋予更高的权重，更加准确地反应用户的当前偏好。将多维因素纳入到推荐模型中，对存在数据稀疏性问题的系统，在忽略个别用户没有对其评分的项目之后，仍可以得到较好的推荐效果。因此，此方法能够有效提高项目推荐系统的准确度。随着推荐系统在电子商务、电子图书馆等领域的广泛应用，基于 TRPT 的项目推荐方法在为顾客购物、读者用书等方面提供完全个性化的决策支持和信息服务的同时，也帮助企业有效地解决信息过载造成的顾客流失问题，为个人和企业创造经济效益。

5 结语

本文首先论述了 web2.0 环境下，在以协同过滤为背景的推荐系统中，考虑标签、项目得分和用户偏好时效性的作用和意义。然后在借鉴协同过滤推荐算法思想的基础上，提出了基于 TRPT 的项目推荐方法，它对提高推荐系统的准确度是有帮助的。最后，对推荐模型的体系结构及其应用前景进行了分析。运用实证方法研究标签、时间和得分对推荐效果的影响程度成为今后探索的方向。

参考文献

1 刘建国,周涛,汪秉宏.个性化推荐系统的研究进展.自然科学进展,2009,19(1):1–15.

2 Yi Z, Li WJ, et al. TagiCoFi: Tag Informed Collaborative filtering. Proc. of the 3rd ACM Conference on Recommender Systems. 2009: 69–76.

3 Yuan C, Guang Q, et al. Model Bloggers' Interests Based on Forgetting Mechanism. Proc. of the 17th International Conference on World Wide Web. 2008: 1129–1130.

（下转第 110 页）

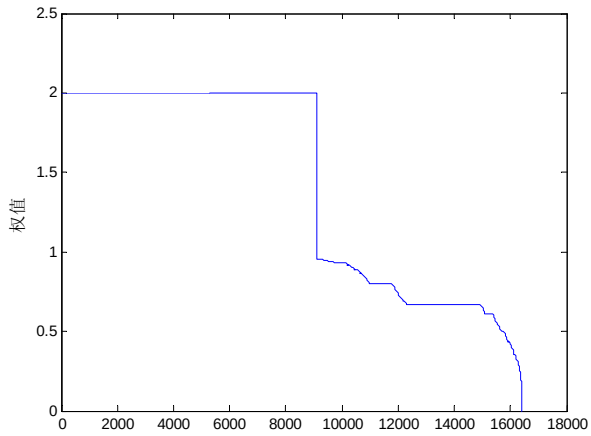

图 4　MI 算法权值分布图

2) 维数不是越高越好，各种算法 5000 维和全维的各项指标几乎相近，低维时 D 算法,DF 算法及综合算法相对有较好的综合特性,MI 算法正确率较高，IG 算法召回率较高。

3) 分析 200 及 100 维的数据，从召回率及正确率这两个评价指标发现：MI 算法容易漏判垃圾邮件，而 IG 算法又容易过滤过度，使用户损失正常邮件。综合方法能够充分利用四种算法的优点，更好的平衡垃圾邮件的检出率和检对率这两个指标。

4) 分析原因，DF 算法保证了取到的特征词是高频词汇；D 算法从类间信息的角度考虑，去除了在两类中概率相近的词语，这就包括常用词，罕见词。IG 算法考虑了单词未出现的情况；而 MI 算法则能够帮助取到一些区分度高的低频特性。取每类的公共集就可以使取到的特征词兼顾到以上各个算法的优点。

6　结语

垃圾邮件用文本矢量表示的最明显的特点就是高维性，针对常用特征提取算法各有所长的特点，本文对各种特征提取算法进行仿真，在分析各种特征提取算法的优缺点的基础上，用一种综合性的特征提取算法，即提取特征词时同时考虑它在每个特征提取算法中的权值排序，取其公共词集构成特征矢量，经分类测试得到了满意的效果。一方面考虑到语料集的限制，对测试肯定会造成一定程度的影响，可以通过加大语料库来提高分类的精度，另一方面 Rocchio 方法本身简单但是受训练文档中的噪声影响较大，同时对某些特征空间非线性可分情况没有处理能力[4]。如果使用 winnow 分类器这类有自修正能力的分类器，相信可以在此基础上进一步提高过滤效果。

参考文献

1 曹麒麟,张千里.垃圾邮件与反垃圾邮件技术.北京:人民邮电出版社,2003.34–56.

2 侯汉清.文本自动标引与自动分类研究.南京:东南大学出版社,2009.57–64.

3 谷波,刘开瑛.中文文本分类中一种简单高效的特征词选择方法.计算机研究与发展,2005,42(增刊):359–360.

4 戴文华.基于遗传算法的文本分类及聚类研究.北京:科学出版社,2008.46–47.

5 王斌,潘文峰.基于内容的垃圾邮件过滤技术综述.中文信息学报,2005,19(5):1–10.

6 李晓飞.垃圾邮件过滤算法研究及系统实现.南京:南京理工大学,2008.

（上接第 205 页）

4 Adomavicius G, Tuzhilin A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans. on Knowledge and Data Engineering, 2005,17(6):734–749.

5 Wei CP, Yang CS, Hsiao HW. A collaborative filtering-based approach to personalized document clustering. Decision Support Systems, 2008,45(3):413–428.

6 Liang HZ, Xu Y, Li YF, et al. Connecting Users and Items with Weighted Tags for Personalized Item Recommendations. Proc. of the 21st ACM Conference on Hypertext and Hypermedia, 2010: 51–60.

7 Zheng N, Li QD. A recommender system based on tag and time information for social tagging systems. Expert Systems with Applications,2011,38(4):4575–4587.

8 Kim HN, Alkhaldi A, Saddik AE, et al. Collaborative user modeling with usergenerated tags for social recommender systems. Expert Systems with Applications, 2011,38(7): 8488–8496.

9 Su JH, Wang BW, Hsiao CY, et al. Personalized rough-set-based recommendation by integrating multiple contents and collaborative information. Information Science, 2010,180(1): 113–131.