情 报 学 报 ISSN 1000 - 0135
第 34 卷 第 5 期 466 - 470, 2015 年 5 月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000 - 0135
Vol. 34 No. 5, 466 - 470 May 2015

doi:10.3772/j.issn.1000-0135.2015.005.003

基于用户兴趣 - 标签的混合推荐方法研究[1)]

李兴华[1] 陈冬林[1] 杨爱民[1] 李 伟[2]

(1. 武汉理工大学电子商务与智能服务研究中心,武汉 430070;
2. 福建新奇特车业服务有限公司,福州 350000)

摘要 传统推荐技术存在冷启动、稀疏性、推荐精度低等问题。其中,可以方便表达用户兴趣偏好的标签推荐存在噪声、一词多义等问题,稳定性较好的用户兴趣刚好可以解决这一问题。然而,在推荐技术领域内,将兴趣与标签相结合的推荐研究相对较少。本文提出基于兴趣 - 标签的推荐算法 ITRA (Interest-Tag Recommendation Algorithm),通过定义计算用户兴趣权重值、用户兴趣相似度、用户候选兴趣集、推荐兴趣 - 标签集、项目推荐集,将该集合作为最终的推荐结果。最后,通过实验证明该算法可以有效的提高推荐结果的准确率。

关键词 个性化推荐 用户兴趣 兴趣标签 电子商务

A Study of Mixed Recommendation Method Based on User Interest-tag

Li Xinghua[1], Chen Donglin[1], Yang Aimin[1] and Li Wei[2]

(1. *Research Center for E-business and Intelligent Services, Wuhan University of Technology, Wuhan* 430070;
2. *Fujian new special vehicle industry service co., Fuzhou* 350000)

Abstract Traditional recommendation technologies have many disadvantages, such as cold start, sparseness as well as low recommendation accuracy. Among these technologies, tag recommendation can express users' interests very well, however it still exists some problems such as noise interference and polysemy. In such situation, users' interests are more stable and can be used to solve the problems mentioned above. While only several studies have combined interest and tags in recommendation area. This paper put forward ITRA (Interest-Tag Recommendation Algorithm) to deal with the condition. ITRA was able to calculate the weight of users' interests and the similarities between users' interests. On this basis, it could get the candidate set of user interest together with the recommendation set of interest-tag, and in the end recommended the items set to users. Finally, the experimental study can verify the improvement of recommendation accuracy by using this algorithm.

Keywords personalized recommendations, user interest, interest tag, e-commerce

1 引 言

推荐系统被认为是解决信息过载最有效的方式,现有推荐技术有:①以计算推荐对象的内容特征和用户模型中兴趣特征二者之间相似性为关键的内容推荐[1]。②目前应用最为广泛,但具有冷启动和数据稀疏性问题的协同过滤推荐算法[2]。③强调关于商品项的领域知识及关于客户的隐式知识中相关推荐规则的基于知识推荐算法[3]。④利用用户

收稿日期:2014 年 7 月 30 日

作者简介:**李兴华**,男,1990 年生,武汉理工大学电子商务研究生(1134099456@qq.com);**陈冬林**,男,1970 年生,博士生导师,主要研究方向:云计算、服务管理、商务智能;**杨爱民**,男,1970 年生,讲师,主要研究方向:智能推荐、企业资源计划(ERP)、企业间供应链集成.

1) 基金项目:国家科技支撑计划(2013BAH13F01)

－项二分图建立关联关系，不考虑客户和项目的属性特征，仅仅将它们看作抽象点的基于网络结构推荐算法[4]，其优点是提高推荐精度、降低了算法复杂性[5]，缺点是进一步恶化数据稀疏性问题[1,2]。⑤基于社交网络的推荐包括：社交网络中基于客户兴趣的实时内容推荐算法[6]；利用客户的网络中心，采用基于客户行为与社交网络结构的推荐方法[7]。⑥通过不同方式结合上述两种或两种以上的方法来改善推荐的性能，以解决基础算法中存在的冷启动和数据稀疏性等问题的混合推荐算法[1,8]。⑦既可以看作是商品内容的萃取，也可以方便用户表达自己的兴趣与偏好的标签推荐，可以部分解决冷启动问题[9]。

其中基于标签推荐常用算法为 user-tag-item 的三部图推荐。在基于标签的项目推荐方法中，张子柯等把标签推荐系统看作是由用户－项目、项目－标签两个二部图组成的三部图，提出了基于标签的扩散推荐算法[10]。在标签推荐系统中，兴趣相似的用户（同一社区成员）很可能使用相似的标签，这对在标签系统中根据用户兴趣进行推荐提供了很好的基础[11]。但由于标签具有噪声、歧义、标签冗余等问题给基于标签的推荐技术研究带来了挑战。然而，用户兴趣的动态性、稳定性和渐变性三个特征反映了用户时间相对长久的信息需求相对稳定的集中在若干信息主题、信息类型、信息源上[12]。有效地避免了标签噪声、歧义、标签冗余等问题。在 Web2.0 下，用户所关注的订阅（如在 Twitter 上）、购买的商品（如在 Amazon 上）、评级（如在视频网站 Netflix 上）、运行的搜索（如在 Google 上）或者某些口味的评论（如在 Hunch 上）对于推荐系统而言都是有价值的个性化信息，对它们进行正确的分析得到用户兴趣及其兴趣偏好程度，再基于用户兴趣的相似性向目标用户进行高效推荐，这在很大程度上解决了传统推荐技术的冷启动和稀疏性问题，更能提供多样性推荐。

现有的推荐理论和方法存在着冷启动、稀疏性、精确度低和多粒度级问题，严重影响了客户体验，不满足大数据环境下面向未来的个性化推荐需求[1,13]。融合客户跨时间、跨系统和跨空间的碎片化知识生成的兴趣图谱及推荐系统被寄予厚望。本文提出基于兴趣－标签的推荐算法 ITRA，旨在结合兴趣与标签优点，改善现有推荐理论中存在的冷启动、稀疏性等问题。

2　基于兴趣－标签的四部推荐机制

文章借鉴了 ODP（Open Directory Project）开放式分类目录搜索系统原理[14,15]。ODP 是互联网最大的分类目录网站，其层次树形目录结构可以帮助用户消除搜索要求的二义性，为用户提供明确的符合其预期的搜索结果，然而，ODP 缺乏用户兴趣偏好权重的表示[16]。因此，本文在借鉴其系统原理定义用户、兴趣、标签及项之间对应关系，同时提出兴趣权重表达用户兴趣的偏好程度。其中用户层为所有用户集合，兴趣层是所有用户的兴趣集合，标签（Tag）将用户兴趣与项目间有机的结合在一起，项即为被推荐的项目实例，所有用户的行为与交互实例组成实例项集合；兴趣权重的值越大，代表某一用户对特定兴趣的偏好程度越高；反之则越低。推荐算法 ITRA 的用户－兴趣－标签－项四部推荐机制图如图 1 所示，主要分为三步：

用户候选兴趣集的计算：根据用户－兴趣矩阵 R（User-Interest Matrix R）中用户兴趣权重计算目标用户与其他用户的兴趣相似度。为降低算法复杂度，设置用户兴趣相似域 w，选取矩阵 S 中除目标用户本身以外相似值大于或等于 w 的用户作为该用户的最近邻，则用户候选兴趣集 CIS（Candidate Interest Set CIS）为目标用户兴趣与最近邻兴趣的并集。

推荐兴趣－标签集的计算：在计算推荐兴趣－标签之前，需要从系统中获取用户的历史浏览数据、用户与项之间的交互行为数据，构建兴趣－标签－项矩阵（Interest-Tag-Item Matrix），兴趣－标签、标签－项之间是一种多对多的关系，因此，利用统计技术分别统计兴趣－标签－项矩阵中每个兴趣下出现在所有项中同一标签的频率，建立兴趣－标签矩阵（Interest-Tag Matrix）。在完成第一步计算后，将用户候选兴趣集与兴趣－标签矩阵（Interest-Tag Matrix）进行一系列计算后得到推荐兴趣－标签集（Interest-Tag Set），为推荐机制的项（Item）推荐提供支持。兴趣－标签是指用标签将用户的兴趣进行标识，便于推荐时检索与分享。

基于用户－兴趣－标签－项四部推荐：完成以上两步推荐的基础操作后，通过推荐兴趣－标签集与标签－项矩阵（Tag-Item Matrix）的映射匹配，选出兴趣偏好最接近用户的项作为推荐结果推荐给目标用户，实现以用户兴趣为基础、以用户为对象的用

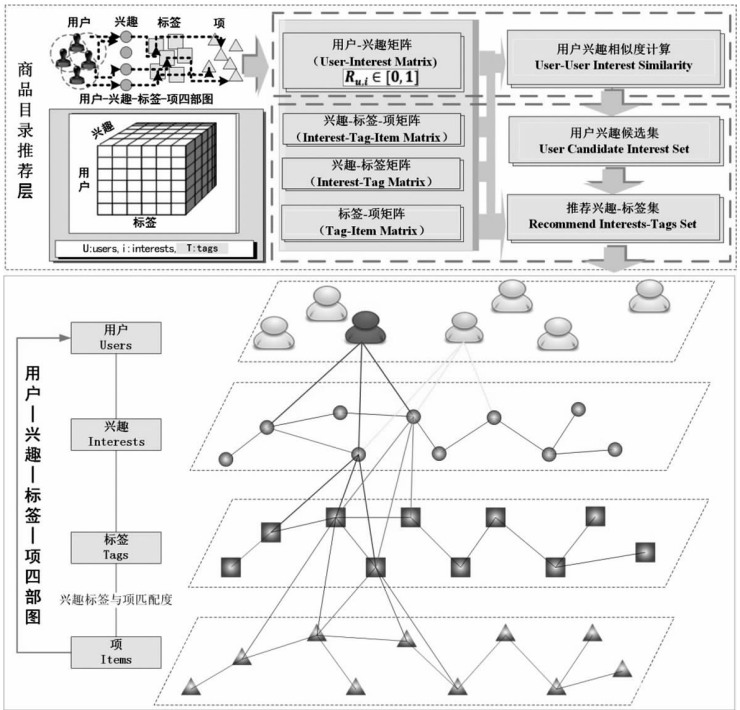

图 1　用户 – 兴趣 – 标签 – 项四部推荐机制图

户 – 兴趣 – 标签 – 项四部推荐机制。

3　基于用户兴趣 – 标签的推荐算法

3.1　算法公式详细说明

不同用户对不同兴趣的喜好程度是有差异的。本文提出兴趣权重的概念来衡量某一用户对特定兴趣的偏好程度。兴趣权重与两个因素有关:①用户与特定兴趣所对应的实例的交互次数;②兴趣权重还与用户的交互频率有关。因此,本文设定用户交互频率来平衡用户的行为习惯对其兴趣权重的影响。定义 $R_{i,p}$ 为用户 u_i 对兴趣 i_p 的兴趣权重,则 $R_{i,p}$ 的计算公式为:

$$R_{i,p} = \frac{N_{u_i}^{i_j}}{F_{u_i}} \tag{1}$$

其中,$N_{u_i}^{i_j}$ 表示用户 u_i 与兴趣 i_j 所对应的实例的交互次数,F_{u_i} 表示用户的交互频率。F_{u_i} 的计算公式为:

$$F_{u_i} = \frac{\sum N_{u_i}}{\sum N} \tag{2}$$

其中,$\sum N_{u_i}$ 表示用户 u_i 的交互总次数,$\frac{\sum N}{n}$ 表示平均每个用户的交互次数($\sum N$ 代表所有用户的交互总次数,n 代表用户总数)。

因此,$R_{(i,p)}$ 的计算公式最终可以表示为:

$$R_{(i,p)} = \frac{N^{i_j} \sum N}{n \sum N_{u_i}} \tag{3}$$

用户候选兴趣集的计算:利用公式(3)计算的用户兴趣权重值构建对应的用户 – 兴趣矩阵 R(User-Interest Matrix R),利用推荐系统中通用的 Pearson 相关系数方法及文献[17 ~ 21]中用户相似性计算方法计算用户兴趣相似度。则:

$$\mathrm{sim}_{(i,j)} = \frac{\sum_{p \in I} R_{(i,p)} * R_{j,p}}{\sqrt{\sum_{p \in I} R_{(i,p)}^2} * \sqrt{\sum_{p \in I} R_{(j,p)}^2}} \tag{4}$$

其中,$\mathrm{sim}_{(i,j)}$ 表示第 i 个用户和第 j 个用户的兴趣相似度,$R_{(i,p)}$ 表示第 i 个用户对第 p 个兴趣的权重值,I 表示用户的兴趣集合。令 l、n 分别为用户集合、兴趣集合元素的数量,经公式(4)计算出两两用户间的兴趣相似度得到一个 $l * n$ 维矩阵——用户兴趣相似矩阵 S,根据值域 w 选取用户最近邻,则其兴趣候选集为用户与最近邻用户的兴趣并集。

推荐兴趣 – 标签集的计算:定义兴趣 – 标签 – 项矩阵 Q(Interest-Tag-Item Matrix Q),分别统计兴趣 – 标签 – 项矩阵中每个兴趣下出现在所有项中同一标签的频率,每个项下出现在所有兴趣中标签的频率,建立兴趣 – 标签矩阵 IT(Interest-Tag Matrix IT),标签 – 项矩阵 M(Interest-Tag Matrix M)。兴趣 – 标签矩阵构建的最终目的是为计算用户兴趣偏

好值,文章参考文献[17～21],定义目标用户对某个兴趣的偏好值等于其对该兴趣的权重加上最近邻兴趣权重与兴趣相似度乘积的和。计算公式如下:

$$X_{(U,I)} = R_{(U,I)} + R_{(O,I)} * \mathrm{sim}_{(U,O)} \quad (5)$$

定义 $X_{(U,I)}$ 表示用户 U 对兴趣 I 的兴趣偏好值,$R_{(U,I)}$ 为用户兴趣权重值,兴趣 $I \in CIS$(CIS 为目标用户候选兴趣集)。

用户兴趣偏好矩阵 X 与矩阵 IT 相乘获取用户推荐兴趣－标签矩阵 A(Recommend Interest-Tag Matrix A),计算公式如下:

$$A_{(I,T)} = X_{(U,I)} * IT_{(I,T)} \quad (6)$$

其中,$A_{(I,T)}$ 表示兴趣 I 下标签 T 的推荐值,$T_{(I,T)}$ 表示兴趣 I 下标签 T 的频率。优先选择 $A_{(I,T)}$ 值比较大的标签作为用户推荐兴趣－标签。

基于用户－兴趣－标签－项四部推荐:定义推荐兴趣－标签集与标签－项矩阵间的乘积得到项目推荐集,计算公式如下:

$$P_{(T,IT)} = A_{(I,T)} * M_{IT_a,t_t} \quad (7)$$

其中,$P_{(T,IT)}$ 表示项 IT 的推荐值,$M_{(IT_a,t_t)}$ 表示第 a 个项对应第 t 个标签的频率。对项目推荐集的值进行从大到小的排序操作。设置推荐值域 y,将推荐值大于或等于 y 的推荐项目推荐给目标用户。

3.2 算法描述

输入:从海量的数据中获取用户对某个兴趣对应项的交互次数 $N_{u_i}^{i_j}$,用户的交互总次数 $\sum N_{u_i}$,所有用户的交互总次数 $\sum N$;各个兴趣下所有项中同一标签的频率 $IT_{(I,T)}$。

Step 1:使用公式(3)计算出各用户的兴趣权重值,记作 $R_{(i,p)}$。

Step 2:获取用户兴趣权重值 $R_{(i,p)}$,使用公式(4)计算用户兴趣相似度 $\mathrm{sim}_{(i,j)}$,筛选用户最近邻,获得目标用户兴趣候选集 CIS。

Step 3:使用公式(5)、公式(6)分别计算兴趣偏好值矩阵、推荐兴趣矩阵,分别记作 $X_{(U,I)}$、$A_{(i,T)}$。

Step 4:使用公式(7)计算项目推荐集,记作 $P_{(T,IT)}$。选取项目推荐值满足值域 y 的项目作为最终的推荐集合。

输出:项目推荐结果。

4 实验结果与分析

4.1 实验数据集与评估准则

实验数据　实验使用的数据集是由 GroupLens 在网站 http://www.grouplens.org 上提供的开放数据集 MovieLens(100K)。该数据集记录了 943 位用户对 1682 部电影的 89 992 条评分记录。该数据集已被广泛应用在推荐系统的研究中,如文献[22]利用 GroupLens 提供的 MovieLens 数据集测试了协同过滤推荐算法的准确性和效率。在该数据集中,电影被分为了 18 个种类,分别是 Action、Adventure、Animation、Children's、Comedy、Crime、Documentary、Drama、Fantasy、Film-Noir、Horror、Musical、Mystery、Romance、Sci-Fi、Thriller、War 和 Western。实验将这些电影分类将被当着用户兴趣类型,而与电影相关的剧情类别、主演姓名等将被当着标签。实验从数据集中选取 60% 作为训练集、40% 作为测试集,这两个集合之间没有交集,但其并集包含了实验数据集中用户和兴趣的所有链接。

评估准则　度量标准的选择是验证个性化推荐算法质量的关键组成部分,好的度量标准能够十分有效地检测出算法的性能及其不足之处。本实验采用 $P@N$(Precision@N, 准确率)作为预测结果的度量标准。$P@N$ 表示生成的 Top-N 推荐列表中,用户喜欢的项目个数与所有被推荐项目的个数 N 的比值,其定义如下[22]:

$$P@N = \frac{\#relevant\ items\ in\ top\ N\ items}{N} \quad (8)$$

4.2 实验结果与分析

为了证明本文提出的方法具有更高的推荐质量,分别计算本位提出的基于用户兴趣－标签推荐算法(方法 1)、基于内容的协同过滤推荐技术(方法 2)、基于标签的推荐技术(方法 3)在不同推荐长度下的 $P@N$ 值,通过比较这三种推荐技术下的 $P@N$ 值分析哪种技术的推荐效果更好。

将随机划分数据集的过程进行 10 次。在指定向用户推荐兴趣长度的情况下(即指定向用户推荐 Top-N 个兴趣),采取区间渐进的办法,分别取推荐兴趣个数为:2,4,…,16,计算三种推荐技术的 $P@N$ 的值。实验结果如图 2 所示。

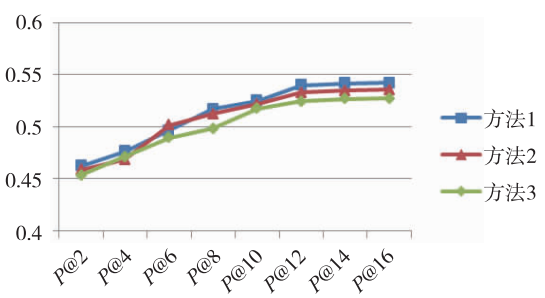

图 2　不同推荐长度下三种方法的推荐准确率

由图2的 $P@N$ 值走势可以看出,本文所提出的方法在准确率上明显优于另外两种方法,并且随着推荐长度的增加,准确率不断增大。当推荐长度大于或等于12时,推荐准确率趋于平衡状态。

5 结 语

随互联网信息数据的快速剧增,如何进行精准化营销与服务成为目前研究的热点之一,推荐技术是解决这一问题的重要手段。本文借鉴了ODP(Open Directory Project)标准,从全网数据信息中提取用户兴趣,这在很大程度上解决了推荐系统中用户的冷启动及稀疏性问题;同时,本文结合用户兴趣推荐与标签推荐方法,提出了基于用户兴趣-标签的推荐算法ITRA,以用户兴趣权重为切入点,通过计算用户兴趣相似度、用户候选兴趣集等计算操作,最终完成项目实例的推荐。经实验证明:该算法与基于内容的协同过滤推荐技术、基于标签的推荐技术相比,能提供更加适当、精准的推荐。然而,由于实验所涉及的兴趣类别较为有限,因此实验还有待进一步的深入。此外,在算法的复杂度上也需要进一步的完善。

参 考 文 献

[1] Bobadilla J, Ortega F, Hernando A, et al. Recommender Systems Survey[J]. Knowledge-Based Systems, 2013(46):109-132.

[2] 杨兴耀,于 炯. 融合奇异性和扩散过程的协同过滤模型[J]. 软件学报,2013,24(8):1868-1884.

[3] Walter Carrer-Neto, María Luisa Hernández-Alcaraz. Social knowledge-based recommender system. Application to the movies domain[J]. Expert Systems with Applications, 2012(39):10990-11000.

[4] 朱郁筱,吕琳媛. 推荐系统评价指标综述[J]. 电子科技大学学报,2012,41(2):163-175.

[5] Li Xin, Chen Hsinchun. Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach [J]. Decision Support Systems, 2013(54):880-890.

[6] Li D, Lv Q, Xie X, et al. Interest-based real-time content recommendation in online social communities [J]. Knowledge-Based Systems, 2012(28): 1-12.

[7] Schall D. Who to follow recommendation in large-scale online development communities [J]. Information and Software Technology, 2014, 56(12):1543-1555.

[8] Joel P Lucas, Nuno Luz. A hybrid recommendation approach for a tourism system[J]. Expert Systems with Applications, 2013(40): 3532-3550.

[9] Zhang Zi-Ke, Zhou Tao, Zhanga Yi-Cheng. Personalized recommendation via integrated diffusion on user item tag [J]. Physica A ,2010(389) 179-186.

[10] Zhang Z K, Zhou T, Zhang Y C. Personalized recommendation via integrated diffusion on user-item-tag tripartite graphs [J]. Physica A: Statistical Mechanics and its Applications, 2010, 389(1): 179-186.

[11] 张富国. 基于标签的个性化项目推荐系统研究综述[J]. 情报学报, 2012, 31(9):963-972.

[12] 杨军,武秀川,郭艳燕. 基于跨系统的个性化搜索系统模型设计[J]. 微处理机,2013(3):41-44.

[13] 汪秉宏,周涛,刘建国. 推荐系统、信息挖掘及基于互联网的信息物理研究[J]. 复杂系统与复杂性科学, 2010,7(2-3):46-49.

[14] Perugini S. Symbolic links in the open directory project [J]. Information Processing & Management, 2008, 44(2): 910-930.

[15] 李建廷. 基于简化 ODP 的用户兴趣模型[J]. 计算机工程与科学, 2010, 32(5): 121-123.

[16] Hyunwoo Kim, Hyoung-Joo Kim. A framework for tag-aware recommender systems [J]. Expert Systems with Applications, 2014(41): 4000-4009.

[17] Lynne Grewe. The interest graph architecture-social modeling and information fusion[C]. Proc. of SPIE 2012, 8392: 1-46.

[18] Lei Li, Li Zheng, Fan Yang. Modeling and broadening temporal user interest in personalized news [J]. Expert Systems with Applications, 2014 (41):3168-3177.

[19] Zhao Z D, Yang Z, Zhang Z K, et al. Emergence of scaling in human-interest dynamics [J], Scientific Reports, 2013(3):3472-3478.

[20] Matias Nicoletti, SilviaSchiaffino, Daniela Godoy. Mining interests for user profiling in electronic conversations [J]. Expert Systems with Applications, 2013 (40): 638-645.

[21] Heung-Nam Kim, Ae-Ttie Ji, Inay Ha, et al. Collaboration filtering based on collaboration tagging for enhancing the quality of recommendation [J]. Electronic Commerce Research and Appllications, 2010(9):73-83.

[22] 谭学清, 何珊. 用户情境下基于信息增益和项目的协同过滤推荐技术研究[J]. 情报杂志, 2014, 33(7):165-170.

(责任编辑 马 兰)