

个性化推荐技术在微课系统中的应用

赵 旭, 吕鹤轩

(大连东软信息学院, 辽宁 大连 116023)

摘 要:随着我国教育信息化、数字化概念的提出,教育信息化资源以极快的速度大量涌现,导致信息呈现爆炸式增长,用户在面临海量信息时,需要花费大量的时间和精力去筛选所需的资源。本文针对此现状提出基于个性化推荐技术实现微课程视频的筛选方案。通过本系统可以为用户提供有针对性的、符合用户兴趣的相关资源,从而使用户获取信息的方式发生翻转,从以往的主动获取变为被动的接收,并保证用户接收的资源对其而言价值最高。

关键词: 推荐系统;协同过滤算法;用户兴趣模型;资源评价机制

中图分类号: TP391 **文献标识码:** A

Application of Personalized Recommendation Technology in Micro-Course System

ZHAO Xu, LV Hexuan

(Dalian Neusoft University of Information, Dalian 116023, China)

Abstract: With the development of educational informatization and digitization in China, the reform of education informatization has made positive progress, but it has also led to explosive growth of information. Users need to spend a lot of time and energy to screen out the resources they need in the massive information. In view of this situation, this paper proposes a corresponding solution to realize video filtering of micro-course based on personalized recommendation technology. Through this system, users can be provided with targeted and relevant resources in line with their interest, so as to change the way users obtain information from the previous active acquisition to passive reception, and ensure that the resources received by users are of the highest value for them.

Keywords: recommendation system; cooperative filtering algorithm; user interest model; resource evaluation mechanism

1 引言(Introduction)

互联网的发展带动了教育的信息化发展,但是我们一方面享受这海量资源带来的便利,另一方面也承受着信息爆炸带来的搜索迷雾问题。那么如何在众多资源中快速、准确地找到自身所需的信息,进而发挥出网络资源的效用,从而最终服务于用户、便利人们的生活是当下互联网领域的重点关注与研究课题,由此越来越多的个性化推荐技术,以及推荐系统应运而生。

本文针对教育领域的微课系统^[1],利用个性化推荐技术可以实现为学生、教师提供有针对性的、符合用户兴趣的相关资源,使用户获取信息的方式发生变革,从以往的主动获取变为被动的接收,并且接收的信息对其最为有用。本文研究的核心主要包括三方面内容^[2]:创建用户兴趣模型;挖掘、表示以及更新用户兴趣项;创建视频资源评价体系;结合个性化推荐算法将用户兴趣与资源特征进行合理匹配,从而完成推荐。

2 用户兴趣模型(User interest model)

用户兴趣模型^[3]是整个系统的数据基础,为后续的推荐功能提供依据和支撑,因此在收集用户数据时,要尽量保证能够较为全方位、无差错的数据收集,良好的用户兴趣收集机制将直接影响到系统的推荐质量,进而直接决定用户在使用系统时的直观感受。

2.1 用户兴趣收集

在本文的研究中,主要通过显式收集和隐式收集两种方式来收集用户数据^[4]。

(1)显式收集

显式收集方式主要包括收集用户自然属性数据以及用户评价反馈数据。其中,用户自然属性数据包括用户所属年级、课程等,这部分数据的主要功能是用用户首次使用本系统时,在系统中无任何用户的个人信息记录,系统可以根据用户自然属性作为用户初始兴趣加入用户兴趣模型中,从而一定程度上避免“冷启动”^[5]问题;用户评价反馈数据是用户在

使用系统时,主动向系统提交他们对资源的评价。

(2)隐式收集

隐式收集的行为主体为系统本身,系统通过技术手段记录用户的操作记录,从本系统自身特点出发,本文中采集用户的搜索、浏览、下载、观看等行为记录作为隐式收集的主要收集数据。采用这种方式的好处是收集过程中,不要用户刻意配合参与,所有的收集行为均发生在服务器端,不会给用户带来任何的使用负担。

通过上述分析,需要寻找一种表达机制,不但能够准确的记录、表达用户行为,同时还要方便系统对行为数据进行分析、计算及统计,通过调研,本文采用日志文件的方式记录表达用户行为。

2.2 用户兴趣表示

针对系统记录的日志文件,我们需要过滤掉噪音信息:即本系统不关注的、对生成用户兴趣没有任何作用的信息,同时尽最大可能提取出对采集用户兴趣发挥作用的数据,最终生成用户的兴趣数据。关于采集的用户兴趣数据,我们更多关注的是用户对哪一类数据感兴趣以及对该类数据的兴趣程度,结合该特点,本文中采用向量空间模型(VSM)^[6]表示法建立和表达用户兴趣。并将用户兴趣模型方法的表示可以表示为以下形式: $\{(tag_1, w_1), (tag_2, w_2), \dots, (tag_i, w_i), \dots, (tag_n, w_n)\}$ 。其中, $tag_i(i \in \{1, 2, 3, \dots, n\})$ 是系统根据收集到的用户行为记录进行清洗、过滤后得到的用户兴趣项, $w_i(i \in \{1, 2, 3, \dots, n\})$ 则代表用户对 tag_i 的感兴趣程度。

同时,需要特殊强调的是冷启动的问题,本文针对第一次使用系统的用户,会将注册用户的自然属性作为其默认的初始兴趣项,并为其赋予初始兴趣权重值,进而解决了系统冷启动时用户兴趣模型数据为空的问题。

模型中的兴趣项权重的计算,本文采用TF-IDF方法^[7]进行计算。TF-IDF的原始计算公式如下:

$$w_{ij} = \frac{tf_{ij} * \log(N/n_i + 0.01)}{\sqrt{\sum_{t_i \in D_j} [tf_{ij} * \log(N/n_i + 0.01)]^2}} \quad (1)$$

其中, w_{ij} 表示特征项 t_i 在文本 D_j 中的权重, tf_{ij} 为特征项 t_i 在当前文档中出现的频次, N 为是训练项目集合中包含的文档数目, n_i 是训练项目集合中包含特征项 t_i 的文档总数。

通过对公式进行分析,我们可以发现用户兴趣的权重值的取值区间为 $[0, 1]$, 权重值的大小与用户对该兴趣项的关注程度高低成正比关系,因此更加证实该方法对于本文切实可行。

2.3 用户兴趣更新

心理学的相关研究发现^[8], 人类大脑中对于某段记忆的深刻程度与该段记忆生成的时间、该段记忆被调取的频率都有着不可分割的关系,间隔时间越长、记忆被调取的频率越低,该段记忆越容易被遗忘,与此同时,新的记忆也不断地生成并存储与我们的脑海之中,因此,在创建用户兴趣模型时,也要考虑人类的这种记忆的逐渐遗忘、不断生成的特点,对用户的兴趣模型数据进行更新。

在本文中,我们侧重的是对于调用记忆的间隔时间这一记忆影响因素对兴趣项权重值的影响,利用的是用户对该兴趣项的最后一次访问时间与当前时间的时间间隔这一数据,将用户原有兴趣项的权重值按如下公式进行衰减处理:

$$\text{newWeight} = (\text{oldWeight})(1 - D\text{value}/Date) \quad (2)$$

其中, oldWeight 、 newWeight 分别是记忆衰退前、后的兴趣项权重值, $D\text{value}$ 是针对该兴趣的访问时间差, $Date$ 为时间调节因子。通过公式可以发现,当访问时间差 $D\text{value}$ 逐渐接近 $Date$ 时,用户对该兴趣项的权重值逐渐降低。

3 微课资源评价机制(Evaluation mechanism for micro-course resources)

本文的目标是系统不但能够为用户生成较为准确的推荐,并且推荐的微课质量为同类型中质量较高、普遍好评的资源,因此本文需要有一套合理的机制能够对微课质量的高低进行评价,而质量的高低最有发言权的则是系统的使用用户,因此本文根据用户的反馈来对资源质量做出评价,用户的反馈行为仍然可以分为显式反馈和隐式反馈两种形式,分别对应着用户评分的两种形式。

用户显式评分:用户主动、直接对视频资源进行的打分行为对应的评分值。本文选用5分制评分机制,定义为 r_1 。

用户隐式评分:隐式评分仍然不需要用户的主观参与,一方面这种收集方式对用户的使用不产生任何负面影响,另一方面这种采集方式采集到的数据相对客观,系统会自动地对用户的观看视频时长进行记录,最终通过用户观看视频时长与视频总时长的比值 p 来反映其对视频的满意程度,计算公式为:

$$p = \begin{cases} t/T_{\text{total}} \times 100\%, & t < T_{\text{total}} \\ 100\%, & T_{\text{total}} \leq t \leq 1.5T_{\text{total}} \\ 0, & t > 1.5T_{\text{total}} \end{cases} \quad (3)$$

其中, t 为观众观看微课视频的实际时长, T_{total} 代表所看微课视频的总时长。隐式评分同样采用五分制的评分制度,则将观众对某视频的隐式评分 r_2 定义为:

$$r_2 = p \times 5 \quad (4)$$

综上,在获取到显式评分和隐式评分后,我们采用两种评分的平均值作为视频的最终总体评分,计算公式如下:

$$r = (r_1 + r_2)/2 \quad (5)$$

4 基于用户的协同过滤算法(User-based collaborative filtering algorithm)

4.1 算法思想

基于用户的协同过滤算法(User-CF)^[9]就是以相似用户为切入点作为推荐依据,将相似用户感兴趣的资源推荐给目标用户。因此,根据算法的思想,其实现主要分为寻找相似用户、用户兴趣度计算、生成推荐三个核心步骤。

4.2 用户相似度计算

如何与目标用户需求接近的用户是本算法需要解决的第一个问题。本文中,因为用户兴趣模型我们采用了向量的表示方式,因此可以计算向量间的夹角余弦值^[10],计算公式分别如下:

$$\text{sim}(u_1, u_2) = \cos(a, b) = \frac{a \cdot b}{\|a\| \times \|b\|} = \frac{\sum_i^m r_{1i} r_{2i}}{\sqrt{\sum_i^m r_{2i}^2} \sqrt{\sum_i^m r_{1i}^2}} \quad (6)$$

其中, a 、 b 分别为用户 u_1 和用户 u_2 的评分向量, r_{1i} 和 r_{2i} 分别表示用户 u_1 和用户 u_2 对视频 i 的评分值。本文研究的研究是利用向量空间模型为计算依据,同时结合基于用户的协同过滤算法,利用上文提到的余弦相似度的计算方法对用户间的相似度进行计算,针对用户 u 和用户 v , 计算二者的相似度,那么

相似公式表达为：

$$\text{sim}_{uv} = \frac{|\text{tag}(u) \cap \text{tag}(v)|}{\sqrt{|\text{tag}(u)| |\text{tag}(v)|}} \quad (7)$$

其中， $\text{tag}(u)$ 和 $\text{tag}(v)$ 表示用户 u 和用户 v 的兴趣集合。通过该方法，系统可以得到与目标用户相似度最大的前 k 个用户集合，设为 $S(u,K)=\{u_1,u_2,\cdots,u_k\}$ 。

4.3 用户兴趣度计算

目标用户最近邻居集合 $S(u,K)$ ，然后利用集合中每个用户与目标用户的相似度，预测目标用户对没有过有效行为视频的兴趣大小，预测公式如下：

$$p(u,j) = \sum_v \text{sim}_{uv} b_{vj}, v \in S(u,K) \cap U(j) \quad (8)$$

其中， $S(u,K)$ 表示和用户 u 的 K 个最近邻居集合， $U(j)$ 表示对 tag_j 有兴趣的用户集合，用户 v 是在用户 u 的最近邻居中对 tag_j 感兴趣的。用户。 sim_{uv} 是用户 u 和用户 v 的兴趣相似度， b_{vj} 表示用户 v 对 tag_j 的兴趣权重。

最终选取用户兴趣集合中兴趣值最大的 M 个兴趣，我们将其定义为 $I_m=\{I_1,I_2,\cdots,I_m\}$ 。

4.4 生成推荐

通过上文中得到的用户兴趣项结合，进而可以在系统中查找到与兴趣项想匹配的相关微课集合作为待推荐资源，不过在待推荐结果最终呈现给用户之前，根据用户的常规使用习惯，我们将用户最有可能关注的视频资源放置在排位靠前的位置进行显示，因此需要对带推荐微课资源的排序进行设计，我们主要考虑两方面因素：用户兴趣值和视频评价，微课资源排序权重计算公式如下：

$$w_v = p(u,j) \times \alpha + R \times \beta, \text{ 其中 } \alpha + \beta = 1 \quad (9)$$

式(9)中， $p(u,j)$ 是表示用户对该类微课的关注程度， R 是该微课的评分值， α 和 β 则分别表示用户关注程度以及视频评价对排序的影响因子。通过计算待推荐微课的 w_v 数值，对所有带推荐微课资源按照 w_v 的值进行降序排序形成最终推荐微课资源，因此最终呈现给用户的将是用户感兴趣且评价较高的视频资源。

5 实验结果(Experiment results)

5.1 实验数据收集

实验时，邀请100位人员进行模拟真实用户测试，实验通过收集测试人员的操作行为数据形成log日志，截取2M的用户行为操作数据对其进行分析提取，并分别收集用户对系统推荐的前三个视频、前五个视频以及前十个视频的满意程度，并形成调查问卷，进而得出实验结果数据。

5.2 实验结果

通过对调查问卷进行统计，得出测试用户对推荐视频资源的满意度统计数据如表1所示。

表1 用户对视频推荐结果满意程度统计表

Tab.1 User satisfaction with video recommendation results

内容	非常满意	满意	不做评论	不满意	非常不满意
Top3视频满意度	27	65	5	2	1
Top5视频满意度	27	62	8	2	1
Top10视频满意度	20	64	13	2	1

通过上述实验结果数据，发现推荐结果能够满足大部分

用户的期望和使用需要。在前十个微课视频推荐中，因为互联网用户的使用习惯，大部分用户在使用时，主要停留在前五条数据中，而不会花费过多精力逐个浏览全部结果，导致不予评论人数增多，进而拉低了用户的满意程度。

6 结论(Conclusion)

本文将推荐技术应用到教育领域的微课系统中，通过采集用户兴趣、生成资源评分，最终将基于用户的推荐算法作桥梁将用户兴趣与微课视频资源链接起来，将用户主动搜索变为被动接收信息，进而完成了个性化的推荐过程。

参考文献(References)

- [1] Yiu-Kai Ng,Jane Linn.CrsRecs:A Personalized Course Recommendation System for College Students[A].Proceedings of 2017 8th International Conference Information,Intelligence, System&Applications[C].Lamaca:IEEE,2017:1-6.
- [2] 冷亚军,黎忠雪.个性化推荐及其相关技术分析[J].内蒙古科技与经济,2019(5):58-60.
- [3] Gong Yuyun,Zhang Qi.Hashtag Recommendation Using Attention-BasedConvolutional Neural Network[A].Proceeding of the 25th International Joint Conference on Artificial Intelligence[C].Palo Alto:AAAI Press,2016:2782-2788.
- [4] 项亮.推荐系统实践[M].北京:人民邮电出版社,2012.
- [5] Rashid A M,Albert I,Cosley D,et al.Getting to know you:learning new user preferences inrecommender systems[C]. Proceedings of the 7th international conference on Intelligent user interfaces.ACM,2002:127-134.
- [6] 邓娟,陈西曲.基于用户兴趣变化的协同过滤推荐算法[J].武汉工业学院学报,2013,32(4):48-51.
- [7] 郭俊霞,徐文生,卢昱.基于用户浏览轨迹的商品推荐[J].计算机科学,2016,43(12):223-228.
- [8] DING Y,LI X.Time weight collaborative filtering[C]. Proceedings of the 14th ACM International Conference on Information and Knowledge Management.ACM,2005:485-492.
- [9] Resnick P,Iacovou N,Suchak M,et al.GroupLens:an open architecture for collaborative filtering of netnews[J].Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work:Chapel Hill,1994:175-186.
- [10] Breese J S,Heckerman D,Kadie C.Empirical Analysis of Predictive Algorithms for Collaborative Filtering[C]. Fourteenth Conference on Uncertainty in Artificial Intelligence,1998:43-52.

作者简介：

赵旭(1989-)，女，硕士，讲师.研究领域：数据挖掘，算法。
吕鹤轩(1996-)，男，本科生.研究领域：数据挖掘，算法。