doi:10.11835/j.issn.1000-582X.2022.201

高校科研资源的个性化融合推荐

刘冬邻
(四川外国语大学 网络信息中心,重庆 400031)

摘要:为满足高校师生对科研资源复杂的个性化服务需求,设计了高校科研资源个性化服务系统,简称个性化科研服务系统(PSRSS,personalized scientific research service system)。全面分析了高校科研用户的个性化科研资源服务需求,设计了基于数据层、融合多种推荐策略的推荐计算层、应用呈现层的多引擎融合推荐系统架构,基于不同推荐场景,比较了不同的推荐算法并对选择的算法进行了针对性优化,探讨了用户模型和科研资源模型的设计,实现了基于资源热度、项目内容相似度、相似用户协同过滤的 Top-N 推荐。系统提升了高校师生获取科研资源的体验,为高校科研资源个性化服务系统建设提供了新思路。

关键词:融合推荐;个性化;推荐系统 ;高校科研
中图分类号:TP520.20 **文献标志码**:A **文章编号**:1000-582X(2022)07-122-09

Personalized fusion recommendation for scientific research resources in universities

LIU Donglin
(Network Information Center, Sichuan International Studies University, Chongqing 400031, P. R. China)

Abstract: To meet the diversified personalized service needs of teachers and students for scientific research resources in universities, a personalized service system for university scientific research resources was designed (PSRSS for short). Firstly, the personalized service needs of users for research resources were comprehensively analyzed. Then, the architecture of multiengine fusion recommendation system was designed with the data layer, the recommendation computing layer which integrates multiple recommendation strategies, and the application presentation layer. Different recommendation algorithms were compared and the selected algorithms were optimized in accordance with the different recommendation scenarios. Next, the user model and scientific research resource model were constructed. Finally, the Top-N recommendation based on the popularity of research resources, similarity of resource content and collaborative filtering of similar users was implemented. The proposed system improves the experience of teachers and students in obtaining scientific research resources and provides new ideas for the development of the personalized service system for scientific research resources in universities.

Keywords: fusion recommendation;personalization;recommendation system;university scientific research

收稿日期:2021-11-13 **网络出版日期**:2022-02-18
基金项目:重庆市教育委员会科学技术研究项目(KJQN201900907)。
Supported by Science and Technology Projects of Chongqing Municipal Education Commission (KJQN201900907).
作者简介:刘冬邻(1979—),男,高级实验师,主要从事数据挖掘、机器学习研究,(E-mail)2843476@qq.com。

高校科研已步入“大数据时代”,各种科研管理系统、科研服务平台收集、储存了海量的科研数据和资源文档[1]。在信息需求越来越个性化的今天,各行业都尝试开发并应用基于各种算法和模型的个性化推荐系统。Amazon 通过在网站上使用推荐系统,对用户的浏览、购买行为进行分析,进而对曾经在该网站有过浏览或购买行为的用户进行个性化推荐。据 VentureBeat 的统计,采用个性化推荐技术,使得亚马逊网站的销售额提高了 30%以上,个性化推荐技术的应用也越来越广[2-3]。

高校师生在科研活动中检索科研资源的时间,占整个科研用时的 50%以上,相较其他类型用户,个性化需求也更多样化、更复杂[4]。目前,师生获取科研数据和科研资源,主要还是使用基于关键字的信息查询检索方式,且国内各类科研管理系统和科研数据服务平台的功能还较单一,无法满足科研用户个性化数据服务需求[5-6]。一方面,科研资源信息过载,面对海量科研数据用户却束手无策,不能方便、快捷地获得需要的科研资源;另一方面,用户要清楚知道自己的资源需求并能明确表示出需求,才能使用搜索引擎查找想要的资源。现有的资源检索或管理系统不能主动把用户可能感兴趣的科研资源推荐给用户,使得宝贵的科研资源得不到充分利用。在大数据背景下,以某外国语大学为例,针对高校师生复杂多样的个性化科研资源需求,探索基于融合推荐的个性化科研资源服务系统的设计。

1　相关工作

1.1　科研资源个性化服务需求

通过对高校科研用户进行问卷调查,总结出师生们主要的个性化资源服务需求为:通过输入自己的研究课题或者论文标题,获得类似的科研项目资源,为自己的科研提供有用的帮助;获得当前本专业相关研究方向的热度值较高的科研资源、科研成果,进而了解当前学科的主要研究方向;了解学科同行当前所从事的研究课题、研究动态,特别是能获得一些自己都不曾想到但又感兴趣的相关资源,为自己的研究找寻参考的方向和可以借鉴的思想。

1.2　融合推荐系统架构

任何单一的推荐策略都不能满足高校用户复杂多样的个性化需求,因此为 PSRSS 设计了融合多种推荐策略的推荐系统架构,由数据层、融合推荐层、应用呈现层组成[7-8]。

数据层:由基础数据和对数据的处理构成。基础数据包括用户信息数据、科研资源数据、用户行为等源数据,科研用户数据主要来自于包含用户个人基本信息的人事系统数据库;科研资源数据主要来自于科研、教改管理系统的用户科研成果数据如科研论文、专著、专利、研究报告、科研项目等;用户行为数据是用户在使用 PSRSS 或其他科研系统时的行为日志数据。数据处理是从业务数据库中抽取所需数据并进行转换、清洗、标准化、融和等预处理,为推荐引擎提供所需数据。

融合推荐层:该层是个性化服务系统的核心,在数据层提供的数据基础上,构建科研用户特征、科研资源项目特征、用户与项目、用户与用户、项目与项目间的关系特征。采用热度推荐、基于 User-CF 推荐和使用 IF-TDF 方法的基于项目内容的推荐算法,构建系统融合推荐引擎,以满足高校科研用户复杂多样的个性化服务需求;该层还包括对系统召回项目进行排序和过滤的模块[9-10]。

应用呈现层:根据应用需要,通过不同的形式向用户呈现推荐的结果。

1.3　科研资源大数据处理

PSRSS 要存储和处理的数据量都是 T 级,同时基于对数据分布式计算和高吞吐量的处理要求,系统采用 Apache 的 Hadoop 大数据技术框架对科研资源大数据进行存储和处理,具体处理过程:

建立数据列表:根据系统需求建立需要的数据列表包括数据的属性、数据之间的关系等。

建立原始数据存储(RDS, raw data stores)和转换后的数据存储(TDS, transformed data strores):物理上通过在 Hive 上建立 2 个数据库来实现,使得所有数据都被分布存储到 HDFS 上。

数据抽取:RDS 作为具体业务系统和 PSRSS 之间的过渡区,它可以避免对源系统的侵入和性能影响,并为细节数据查询提供支持。使用 Sqoop 把各业务系统相关数据抽取到 RDS,使用 Flume 从日志文件中获取用户从外网使用科研资源的数据。

数据转换与装载:建立数据列表到 RDS 的映射,根据融合推荐系统的需要,使用 HiveQL 脚本对数据进

行转换和处理，包括对数据进行去重、补全、查错纠错、标准化等处理，将数据从 RDS 装载到 TDS 中。完成首次的数据抽取、转换、装载(ETL，Extract、Transform、Load)过程后，还需要根据系统需要定期执行数据 ETL 过程，比如按照每天进行一次自动化的增量数据 ETL 过程。

2 科研用户模型和资源项目模型的构建

建立科研用户和科研资源项目之间的关联，实现个性化推荐服务，推荐系统要经过构建科研用户模型和科研资源模型、根据用户特征运用不同的推荐算法对资源项目进行召回计算、向用户呈现科研资源推荐列表这 3 个重要步骤[11-12]。科研用户模型和科研资源模型决定着 PSRSS 的输出。

2.1 构建科研用户模型

在 PSRSS 的用户模型中包括用户基本信息和用户的资源兴趣模型，即用户在使用系统和资源时的一些行为信息，见表 1 所示。

表 1 个性化科研服务系统的用户大数据

Table 1 User big data of PSRSS

数据维度	数据类别	主要内容	数据来源
背景信息数据	基本属性数据	工号、性别、专业、职称、研究方向等	人事、教务等管理系统
背景信息数据	科研信息数据	项目、论文、专著、专利、学术报告等	科研管理系统等
用户行为数据	用户认证数据	账户注册、登录、设备等	人事、认证及 PSRSS 等
用户行为数据	科研资源使用	资源浏览、点赞、收藏、下载、标签设置	资源数据库、PSRSS 等

PSRSS 要向用户推荐他们感兴趣的各种科研资源，不仅要记录用户对资源项目的具体行为数据，还要记录用户使用 PSRSS 的行为数据，如用户浏览某个资源项目内容的具体时长，这些行为数据将用于项目热度和用户兴趣模型的更新。

由于高校科研用户在使用个性化科研服务系统时，往往只专注于所需资源项目的内容本身，他们会查询、浏览阅读、下载获取，不太会对相应项目进行主动评价，很难获得用户对资源项目的显式行为记录。因此，采用隐式的方式，记录并利用用户使用 PSRSS 和资源数据库的行为日志，建立并更新用户模型。表 2 为用户对科研资源行为描述。

表 2 用户对科研资源行为描述

Table 2 Description of users' behavior towards scientific research resources

用户行为	类型	特征	描述
收藏	隐式	取值为 0 或 1	该行为意味着用户已经产生一定兴趣
阅读	隐式	取值为 0 或 1	用户阅读时长反映用户的兴趣度，有噪音
下载	隐式	取值为 0 或 1	基本得到用户的认可

本系统用户模型使用 20 世纪 70 年代由 Gerald Salton 等提出的 VSM(向量空间模型，Vector Space Model)表示，该模型最初用于处理文档，通过识别并获取文档的 N 个关键字特征以表示文档，分别为每个特征赋予合适的权值 W，进而构造一个表示该文档的特征向量。当文档被表示为文档空间的向量后，就能计算不同文档向量间的相似度并据此度量文档间的相似性。在 PSRSS 中，当用户对某个科研资源项目做出某种行为时，其行为值为 1，这些行为反映了用户对资源项目的不同兴趣度，赋予每种行为不同的权值，$W_1 \sim W_3$ 的取值为 0～1 且权值总和为 1。用户对科研资源项目的特征行为如表 3 所示。

表 3　用户对科研资源行为示例

Table 3　Examples of user behavior towards scientific research resources

资源项	收藏（权值 W_1）	阅读（权值 W_2）	下载（权值 W_3）	兴趣度 P
item1	1	1	0	$p_1=0.6$
item2	0	0	1	$p_2=0.4$
Item3	0	1	0	$p_3=0.4$
Item4	1	0	1	$p_4=0.6$
Item5	0	1	1	$P_5=0.8$

根据用户使用 PSRSS 的行为,建立用户偏好(UP,user preference)模型,模型表达式为

$$U\{(r_1,p_1),(r_2,p_2),\cdots,(r_j,p_j)\},\tag{1}$$

式中:r_j 是编号为 j 的科研资源项目;p_j 是用户对 r_j 的综合行为偏好值[13-14],可根据表 3 设定的相应行为的权值计算得到。

2.2　构建科研资源模型

采用基于资源项目内容主题模型的资源推荐策略,通过给能反映资源项目主要内容特征的主题计算权值向量,从而使用该向量计算得到资源项目间的相似度,可以比较精确地向用户推荐其可能感兴趣的科研资源。

高校的主要科研资源类型有科研论文、研究报告、著作、纵向和横向课题、专利、各种比赛成果等。考虑到进行基于科研资源项目内容推荐的需要,特别是推荐算法中引入项目时间因素的改进设计,在对科研资源进行建模时设计了包括资源项目 ID、资源建立时间、资源长度、资源类型、资源内容关键字特征作为科研资源模型的元素,模型表达式为

$$R\{r_{\mathrm{ID}},r_{\mathrm{type}},r_{\mathrm{duration}},r_{\mathrm{length}},k[i]\},\tag{2}$$

式中:

1)r_{type}为科研资源项目的类型,在系统冷启动阶段向用户做热度推荐时,将根据用户的专业、研究方向、资源类型提供初始的资源项目推荐,比如向英语学院研究英语国家文化的教师推荐相应类型的论文或其他资源。

2)r_{duration}是资源项目在系统中发布后存在的时间,反应了资源项目的新旧程度,在推荐过程中,我们要考虑时间因素对用户兴趣度的影响。

3)r_{length}为资源项目内容长度,目前主要的资源项目,其内容形式主要以文本为主,用户阅读浏览的时间和项目内容的长度,共同决定了用户对该资源项目的兴趣度。

4)$k[i]$为资源项目标题关键字列表,在进行基于项目内容的推荐时,使用 TF-IDF 方法从资源标题计算得到该资源项目的关键字列表。

3　算法选择与优化

3.1　算法选择

个性化推荐算法是个性化科研服务的基础,主要的推荐算法有基于内容(content-based)的推荐、基于协同过滤(collaborative filtering)的推荐、基于关联规则(association rule-based)的推荐、基于效用(utility-based)的推荐、基于知识(knowledge-based)的推荐和组合(hybrid)推荐等。

基于内容的推荐是在项目内容信息上做出推荐,不需要用户对项目进行显式评价操作,可通过使用机器学习的方法从描述内容特征的事项中,获取用户的兴趣特征并找到与用户感兴趣的相似内容向用户推荐,可通过增加特征维度的方法来提高该算法的推荐精度。基于内容的推荐,不需要大量的用户-项目评分记录,

可用于新建立的资源项目的推荐，解决项目冷启动问题。

协同过滤推荐算法包括基于用户的协同过滤（User-CF）和基于项目的协同过滤（Item-CF），是一种基于近邻的推荐算法[15]。在电商商品和图书馆资源推荐时多采用 Item-CF，因为用户在找寻这类物品时的兴趣是比较稳定的，因此可以向其推荐与当前浏览物品相似的商品。在 PSRSS 中，当要向用户推荐有关本专业的同行当前关注的科研资源时，科研资源的时效性、专业性和热度，比根据用户的系统使用日志学习得到的兴趣更有用。基于用户的协同过滤推荐策略还能向用户推荐可能让其惊喜的资源项目。

根据 PSRSS 的应用场景需要，融合推荐引擎在系统冷启动阶段，采用基于项目热度的推荐算法，向用户推荐相关专业和研究方向的各类热度值较高的科研资源；在用户有了较多的系统使用行为记录后，选用基于用户的协同过滤推荐算法，向用户推荐有相似兴趣的本学科专业的同行感兴趣的科研资源；当用户收藏、阅读或下载了某项科研资源时，选用基于内容的推荐算法，向用户推荐与其当前感兴趣的资源相似的科研资源。

3.2 算法优化

3.2.1 项目热度值计算

用户刚开始使用 PSRSS 时，系统是无法向用户提供个性化服务的，面临用户冷启动问题，此时采用基于项目热度的推荐算法，对科研资源基于专业学科、研究方向等基本信息进行划分，然后按照项目热度对科研资源进行排序，将热度值较高项目推荐给感兴趣的用户。

当一个资源项目录入系统数据库后，就为其初始化一个热度分 H_0，项目也就同时进入了推荐候选列表，不同科研资源的初始热度分是不一样的，可以根据资源类别并按照作者的专业水平如专业职称等条件，赋予不同资源不同的初始热度值。随着资源项目不断被用户阅读、收藏、下载，对应地被用户行为影响的热度 H_1 不断增加。还有影响资源热度的其他因素 H_2，他们会使资源热度降低，比如时间因素。常用项目热度值公式为

$$S = H_0 + H_1 - H_2。 \tag{3}$$

考虑科研项目热度随时间呈指数增长的衰减趋势，可采用结合牛顿冷却定律改进后的项目热度计算公式，来计算资源项目的热度，为

$$S = \frac{(H_0 + H_1)}{H_2}, H_2 = \exp(k \times (T_1 - T_0)), \tag{4}$$

式中：H_0 为项目初始热度分；H_1 为用户行为带来的热度增加；H_2 为随时间衰减的热度；$T_1 - T_0$ 为项目资源发布至今的时间长度；k 为冷却系数，根据具体应用实验调整。

3.2.2 项目向量化

PSRSS 的主要推荐内容是非结构化的科研资源文档，不能直接将其映射到向量空间，这些资源的标题包含了关于该资源的核心关键信息，能反映资源的主要内容特征，用户也主要是利用各个资源项目的标题信息来对下一步的动作如点开阅读、收藏、下载或者直接略过，做出决定的。利用 TF-IDF 算法从项目标题提取出项目关键词，将关键词的 TF-IDF 值作为该关键词的权值，将包含项目核心特征信息的项目标题映射为表示项目的特征向量，用以计算项目之间的相似度[16-17]，进行基于内容的推荐。

F_{TF} 为词条在文档中的出现频率，词条 j 在文档 F_i 中的出现频率为

$$F_{TF} = \frac{C_j}{|F_i|}, \tag{5}$$

式中：C_j 为词条 j 在文档 F_i 中出现的次数；$|F_i|$ 为文档 F_i 中全部词条的数目。

F_{IDF} 为词条在文档集合中的区分能力，对于一个新建的资源项目，词条 j 可能在其他文档中一次也没有出现，因此采用进行了平滑处理的 F_{IDF} 为

$$F_{IDF} = \log \frac{N}{1 + \sum_{i=1}^{N} I(j, F_i)}, \tag{6}$$

式中：N 为所有文档总数；$I(j, F_i)$ 是表示文档 F_i 是否包含词条 j 的指示函数，若包含则为 1，不包含则为 0。

词条 j 在文档 F_i 中的 $F_{\text{TF-IDF}}$ 值为

$$F_{\text{TF-IDF}} = F_{\text{TF}} \times F_{\text{IDF}}。\tag{7}$$

3.2.3　User-CF 推荐的优化

User-CF 推荐算法的核心是要构建高校科研用户和科研资源项目的关系矩阵。根据用户是否点击浏览或收藏、下载资源文档等行为构建用户特征向量,使用综合用户偏好值 p_i 作为矩阵项的值,建立用户和项目关系矩阵,并用于计算用户相似度[18]。

1)时间因素。

用户在点击浏览项目内容时,阅读时间的长短反映了其对项目的兴趣程度,打开资源一掠而过还是花时间仔细阅读,反映出对项目截然不同的兴趣程度。使用阅读时间影响参数 λ 对用户阅读行为的评分进行修正,将用户 u 阅读某个资源项目 i 的时间表示为 $D_{u,i}$,项目内容长度为 S_i,用户阅读过的该类资源的平均长度为 S_u,用户阅读该类资源的平均阅读耗时为 D_u,α 为常系数,可通过实验取得合适数值。得到阅读时间影响参数 λ 为

$$\lambda = \exp(\alpha(D_{u,i} - S_i \times (D_u / S_u)))。\tag{8}$$

2)热点影响。

知名专家、教授的科研项目和成果会成为大家都关注的热门资源,但并不意味着这些用户之间就有共同的兴趣,因此采用对热门项目进行惩罚的兴趣相似度为

$$W_{u,v} = \frac{\sum_{i \in N(u) \cap N(v)} \frac{1}{\lg(1 + |N(i)|)}}{\sqrt{|N(u)||N(v)|}},\tag{9}$$

式中:$N(u)$、$N(v)$分别表示用户 u 和用户 v 浏览或下载过的科研资源集合;$N(i)$是对资源项目 i 有过行为的用户集合,资源 i 越热门,$N(i)$就越大。

3)稀疏数据的计算。

通常不同院系、专业的科研用户之间并没有什么交集,所以建立的用户项目关系矩阵是一个稀疏矩阵,计算用户相似度时,很多数据没有必要计算。可通过建立项目到用户的倒查表,进行如下优化:

根据用户行为表数据,建立项目到用户的倒查表 M,表示该项目被哪些用户产生过行为。

根据倒查表 M,建立用户相似度矩阵 $\boldsymbol{H}$。在 M 中,对每个项目 i,设其对应的用户为 a、b,如果用户 a、b 同时对项目 i 产生过行为,在 $\boldsymbol{H}$ 中更新对应位置的元素,$\boldsymbol{H}[a,b] = \boldsymbol{H}[a,b] + 1$,$\boldsymbol{H}[b,a] = \boldsymbol{H}[b,a] + 1$。这样扫描完一次倒查表 M 之后,就能计算得到完整的用户相似度矩阵 $\boldsymbol{H}$。

4　科研资源的 Top-N 推荐

这个阶段就是在优化根据应用场景需要选择的推荐算法基础上,计算用户对还没有使用过的科研资源的兴趣度,基于用户兴趣度和其他的资源特征,对待推荐资源列表按降序进行排序,将列表前面的 N 项资源推荐给用户。

4.1　用户冷启动阶段

这个阶段,根据项目的热度值为用户进行推荐,使用式(4)计算项目热度值。可以根据作者专业职称级别,为不同用户设置不同的权值如:中级及以下作者权值为 0.6,副高级作者权值为 0.8,正高级及以上作者权值为 1。根据 $H_1 = 0.2 \times$ 收藏次数 $+ 0.4 \times$ 阅读次数 $\times \lambda + 0.4 \times$ 下载次数,计算用户行为对项目 H_1 分值的更新。系统启动阶段,可以综合考虑作者特征和资源特征为每类资源赋予不同的初始热度值,系统运行后,可以结合每类资源的平均热度值计算新建项目初始热度值 H_0。在此基础上,结合项目作者的权值使用式(4)便可计算出每个资源项目的当前热度值,根据资源类别对每类资源按热度降序排序,将与用户专业和研究方向相关的排名靠前的 N 项各类资源推荐给用户。

4.2　用户协同过滤推荐

根据用户使用科研资源项目产生的用户行为数据,构建项目用户行为倒查表,然后利用式(9)计算用户相似度矩阵。

接下来找到和目标用户最相近的 K 个用户，同时找到他们喜欢的而用户还没有使用过的科研资源项目，根据用户兴趣度模型得到用户对未使用过的资源的兴趣度

$$P_{u,i}=\frac{\sum_{k}^{n}(W_{u,k}\cdot R_{k,i})}{\sum_{k}^{n}W_{u,k}}, \tag{10}$$

式中：n 为和用户 u 相似的用户总数；$W_{u,k}$ 为利用式(9)计算得到的用户 u 和用户 k 的相似度；$R_{k,i}$ 为用户 k 对项目 i 的综合评分，是使用时间影响参数进行修正，计算得到的用户对项目综合评分；$P_{u,i}$ 为预估用户 u 对项目 i 的评分。根据对候选项目按用户兴趣度值降序排序的结果，将前面 N 项资源推荐给用户。

4.3 相似资源项目推荐

文中使用 Python 的 jieba 库作为分词工具，对资源文档标题进行分词处理，在此基础上去除停用词，然后使用 TF-IDF 方法计算单词的 TF-IDF 值，构造项目标题关键词向量。

当用户对某个资源项目进行了阅读、下载等感兴趣的操作，系统便根据当前项目的关键词向量，使用余弦相似度公式(11)，计算其与其他该类项目的相似度，然后依据按项目相似度降序排序的结果，向用户做 Top-N 推荐。

$$\mathrm{sim}(q,s)=\frac{\sum_{i=1}^{n}q_i\times s_i}{\sqrt{\sum_{i=1}^{n}(q_i)^2}\times\sqrt{\sum_{i=1}^{n}(s_i)^2}}, \tag{11}$$

式中：$\mathrm{sim}(q,s)$ 为示资源项目 q 和 s 的相似度；q_i 为资源项目 q 的第 i 个特征；s_i 为资源项目 s 的第 i 个特征；n 为资源项目特征向量的维度。由于 PSRSS 中的科研资源项目一般会进行集中更新，因此可以采取一定策略定期离线计算项目标题特征向量，甚至提前计算每类科研资源项目间的内容相似度，以提高系统运行效率。

5 系统效果评估

针对系统的融合推荐引擎，使用推荐准确率作为评价系统推荐效果的评估指标，主要以用户使用 PSRSS 系统产生并存储在用户资源项目评分表 user_res_items_score 数据表的数据作为实验数据见表 4 所示，这些数据是用户对科研资源的各种操作记录如内容浏览、下载、收藏。该表有用户数 1 206，资源项目数 8 124，表项即用户对资源项目操作数 35 215，将科研资源数据的 80%用作训练集，20%用作测试集并计算系统融合推荐引擎的推荐准确率。

表 4 用户资源项目评分表的数据示例

Table 4 Data example of table user_res_items_score

UserID	resID	Tag	day	mon	year	hour	minu	sec
990013××	01201011	2	5	9	2021	12	23	3
990013××	03120045	1	6	9	2021	11	23	22
990013××	12004536	3	6	9	2021	12	23	57
990006××	03110187	1	15	7	2020	09	34	12
990006××	03110698	2	12	6	2019	10	13	1
990006××	03110427	3	21	11	2019	11	05	45

针对基于项目热度和基于项目内容的推荐，分别计算了推荐列表长度 N 为 5，8，10，12，15 的推荐准确率，如图 1 所示。结果显示，在推荐列表长度 N 为 5 时有较好准确率，随着 N 的增大，准确率逐渐下降。当

N 较小时，基于项目热度的推荐效果更好，这反映出科研用户对当前热点科研项目的关注度较高。当 N 继续增大后，基于内容的推荐效果更好，反映出此时科研的学科专业性及用户对与自己当前研究内容相关的科研资源的关注度，对推荐效果有更大的影响。

针对基于 User-CF 的推荐，分别计算了相似用户数 K 为 3、5、8、10、15，N 为 5 时的推荐准确率如图 2 所示，随着近邻的增加，推荐准确率有较明显改善，在 K 为 8 时最好，之后开始下降，反映出由于科研的专业性，能帮助有效协同过滤的用户数是有限的。

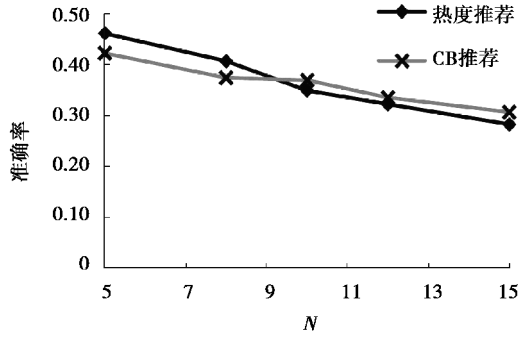

图 1　基于热度与内容相似度推荐

Fig. 1　Recommendation based on popularity and content similarity

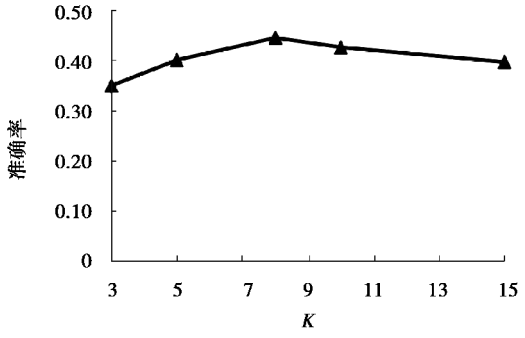

图 2　基于 User-CF 的推荐

Fig. 2　Recommendation based on User-CF

6　结束语

文中调研了高校科研用户的科研资源个性化服务需求，设计了融合推荐系统架构。根据应用场景选择合适的推荐算法并进行了针对性的优化，考虑阅读时间长短对用户兴趣度的影响，加入阅读时间影响因子以修正用户兴趣度值的计算；建立资源项目到用户的倒查表，解决稀疏数据的计算效率问题；在进行基于内容的推荐时，利用科研用户的专业、研究方向等特征进行分类、排序，提高推荐的准确性；利用用户权值和时间影响因子计算项目热度值，并解决了系统冷启动问题。结合多种推荐策略，构建了融合推荐引擎，提高了推荐效率和推荐准确率，为个性化科研资源服务系统的建设提供了新的参考。

本研究还可进一步挖掘高校科研用户的大数据资源服务需求，优化系统架构，提高用户推荐满意度；为其他系统应用设计 API 接口，拓展向师生主动推荐科研资源的渠道。

参考文献：

[1] 覃福钿，李晶. 大数据对高校教学研的影响与探索[J]. 计算机工程与科学，2019，41(S1)：238-241.
Qin F D，Li J. Influence and exploration of big data on university teaching and research[J]. Computer Engineering & Science，2019，41(S1)：238-241.(in Chinese)

[2] Linden G，Smith B，York J. Amazon.com recommendations：item-to-item collaborative filtering[J]. IEEE Internet Computing，2003，7(1)：76-80.

[3] Gomez-Uribe C A，Hunt N. The netflix recommender system：algorithms，business value，and innovation[J]. ACM Transactions on Management Information Systems，2016，6(4)：1-19.

[4] 陈媛媛. 高校科研数据管理服务能力研究[J]. 情报杂志，2020，39(6)：203-207.
Chen Y Y. On research data management service ability of colleges and universities[J]. Journal of Intelligence，2020，39(6)：203-207.(in Chinese)

[5] 刘兹恒，曾丽莹. 我国高校科研数据管理与共享平台调研与比较分析[J]. 情报资料工作，2017(6)：90-95.
Liu Z H，Zeng L Y. Investigation and comparative analysis of scientific research data management and sharing platform of universities in China[J]. Information and Documentation Services，2017(6)：90-95.(in Chinese)

[6] 王欣，张冬梅，闫凤云，等. 大数据环境下基于科研用户小数据的图书馆个性化科研服务研究[J]. 情报理论与实践，2017，40(10)：85-90，95.

Wang X, Zhang D M, Yan F Y, et al. Research on the personalized scientific research service based on small data of scientific research users in library under the big data environment[J]. Information Studies: Theory & Application, 2017, 40(10): 85-90, 95.(in Chinese)

[7] 冀振燕，吴梦丹，杨春，等. 可扩展的融合多源异构数据的推荐模型[J]. 北京邮电大学学报，2021，44(3)：106-111.
Ji Z Y, Wu M D, Yang C, et al. Scalable recommendation models fusing multi-source heterogeneous data[J]. Journal of Beijing University of Posts and Telecommunications, 2021, 44(3): 106-111.(in Chinese)

[8] 刘冬邻. 基于软件复用的外语听说考试系统设计与实现[J]. 西南师范大学学报(自然科学版)，2016，41(6)：125-131.
Liu D L. On design and implementation of foreign language listening and speaking web test system based on software reuse[J]. Journal of Southwest China Normal University (Natural Science Edition), 2016, 41(6): 125-131.(in Chinese)

[9] 李稳安，陈柳柳，陈实. 基于注意力模型的多模态特征融合雷达知识推荐[J]. 重庆大学学报，2021，44(7)：34-42.
Li W A, Chen L L, Chen S. A multi-modal feature fusion radar knowledge recommendation method based on attention mode[J]. Journal of Chongqing University, 2021, 44(7): 34-42.(in Chinese)

[10] Khalid O, Khan S U, Zomaya A Y. Big data recommender systems-volume 1: algorithms, architectures, big data, security and trust[M]. Institution of Engineering and Technology, 2019.

[11] Parmezan A R S, Lee H D, Spolaôr N, et al. Automatic recommendation of feature selection algorithms based on dataset characteristics[J]. Expert Systems With Applications, 2021, 185: 115589.

[12] Fang X J, Wang J Y, Seng D W, et al. Recommendation algorithm combining ratings and comments[J]. Alexandria Engineering Journal, 2021, 60(6): 5009-5018.

[13] Rajendran D P D, Sundarraj R P. Using topic models with browsing history in hybrid collaborative filtering recommender system: experiments with user ratings[J]. International Journal of Information Management Data Insights, 2021, 1(2): 100027.

[14] Bertani R M, Bianchi R, Costa A R. Combining novelty and popularity on personalised recommendations via user profile learning[J]. Expert Systems with Applications, 2020, 146: 113149.

[15] Pérez-Almaguer Y, Yera R, Alzahrani A A, et al. Content-based group recommender systems: a general taxonomy and further improvements[J]. Expert Systems with Applications, 2021, 184: 115444.

[16] Ni J J, Cai Y, Tang G Y, et al. Collaborative filtering recommendation algorithm based on TF-IDF and user characteristics[J]. Applied Sciences, 2021, 11(20): 9554.

[17] Chen J, Wang X S, Zhao S, et al. Deep attention user-based collaborative filtering for recommendation[J]. Neurocomputing, 2020, 383: 57-68.

[18] 石鸿瑗，孙天昊，李双庆，等. 融合时间和类型特征加权的矩阵分解推荐算法[J]. 重庆大学学报，2019，42(1)：79-87.
Shi H Y, Sun T H, Li S Q, et al. A matrix factorization recommendation algorithm with time and type weight[J]. Journal of Chongqing University, 2019, 42(1): 79-87.(in Chinese)

（编辑 詹燕平）