第50卷第8期 中国科学技术大学学报 Vol. 50, No. 8
2020年8月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Aug. 2020
文章编号:0253-2778(2020)08-1072-05

基于知识推荐的校园百科平台研究

任 敏,许 玲,王 峰,吴 超

(中国科学技术大学网络信息中心,安徽合肥 230026)

摘要:2018年中国科学技术大学推出"校园百科"项目,旨在利用智能化技术实现校园文化积累与推广的新途径."校园百科"的推出是以建设校园文化积累的知识库和校园文化分享平台为目的,用于鼓励师生积极参与校园文化建设,并为师生提供文化信息智能化检索和个性推荐的服务.为此以中国科学技术大学校园百科建设为背景,重点介绍了校园百科平台建设中所涉及的关键技术(层次多标签分类、智能搜索和协同过滤标签推荐),并展示了校园百科平台的设计架构和主要功能,最后简要介绍了我校校园百科的使用评估.

关键词:校园百科;层次多标签分类;全文检索;协同过滤标签推荐
中图分类号:TP391.3 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2020.08.005

引用格式:任敏,许玲,王峰,等. 基于知识推荐的校园百科平台研究[J]. 中国科学技术大学学报,2020,50(8):1072-1076.
REN Min, XU Ling, WANG Feng, et al. Campus encyclopedia platform based on knowledge recommendation[J]. Journal of University of Science and Technology of China, 2020,50(8):1072-1076.

Campus encyclopedia platform based on knowledge recommendation

REN Min, XU Ling, WANG Feng, WU Chao

(*Network and Information Center, University of Science and Technology of China, Hefei* 230026, *China*)

Abstract: In 2018, University of Science and Technology of China (USTC) released its encyclopedia platform of the university, which provides intelligent and digital means to accumulate and disperse cultural knowledge of the university. The release of Encyclopedia Platform aims to build up a knowledge base of the campus culture, encouraging the staff and students to actively participate in the recording and sharing of the unique campus culture of USTC, and provides intelligent retrieval and knowledge recommendation services. We present a thorough analysis of the key intelligent technologies of the encyclopedia platform, i. e., hierarchical multi-label classification, intelligent retrieval technologies, tags and collaborative filtering based recommendations, and introduce the architecture and the key functions of the platform. Finally, we briefly discuss its usage statistics.

Key words: campus encyclopedia, hierarchical multi-label classification, full-text search, tag-based collaborative filtering recommendation

0 引言

随着"互联网+教育"概念的深入推广,教育教学信息化改革全面展开,但蕴含丰富校园知识和体现校园价值观的校园文化的智能化推进没有得到足够的重视.校园文化是学校长期发展积淀形成的共识价值体系,也是可记录、可存储、可衡量的宝贵数据资源,因此建设校园百科知识库,推进数字化校园文化;为师生提供平等的、共享参与的校园知识协作编辑平台,拓展传播和学习渠道;提供知识推荐和智能检索服务,成为校园百科智能化建设的目标.

众所周知,维基百科(Wikipedia,Wiki)是全球最大且最受欢迎的网络百科项目.它以多人协作编写的方式实现知识的快速整合,形成知识网络系统,并通过强大的检索功能实现信息共享和流通,具有访问的便捷性、自组织性、可增长性和开放性的特点.2006年,我国知名的中文搜索引擎公司百度推出了百度百科,它是一个涵盖各领域知识的中文信息收集平台.截至2020年5月底,百度百科已收入超过1747万词条,参与编辑的网民超过了711万人.2020年4月30日,头条百科测试版上线,作为知识类搜索产品,其不仅涵盖传统百科的内容,而且将时代热点和动态事件纳入了词条体系.知识网络和知识搜索一直备受行业巨头的关注,"碎片化学习"和"个性化学习"的概念深入人心,因此引

收稿日期:2020-06-05;**修回日期**:2020-08-21
作者简介:任敏(通讯作者),女,1983年生,硕士/工程师.研究方向:软件开发与管理,智能化软件研究与推广.
Email:renmin66@ustc.edu.cn

入百科的概念为校园文化建设所用，建设面向校园文化领域的知识库和个性化知识检索与推荐平台，用于积累和传播校园文化知识，对智慧校园建设有重要的实践意义.

1　关键技术

2018 年，中国科学技术大学推出“校园百科”平台，它作为新一代校园文化传播与知识共享的智能化系统，融合了当前主流的人工智能技术和大数据技术. 本节将从词条组织方式、智能搜索和个性化推荐 3 个方面介绍本平台涉及的关键技术.

1.1　词条组织方式

词条分类是校园百科平台最重要的问题之一. 树状图具有明确的层次结构且支持高效的信息检索技术，长久以来被广泛用于解决各种分类问题. 传统的分类法具有类别互斥的特性，这对具有交叉概念的词条，难以实现精确分类. 例如，词条“郭沫若”既属于“社科人物”类别，同时也属于“校园人物”类别，很难将其归类于任何一类.

为解决上述问题，维基百科的词条组织采用有向无环层次结构作为词条分类体系，并结合主题法对词条进行分类管理，允许子类可以有多个父类，词条具有单向多对多的从属关系，从而避免了传统分类法类别互斥的局限性，解决了词条类别从属多个类别的问题，如图 1 所示[1]. 对于用户来说，这种多层次类别命名也会产生新的问题. 如“计算机应用”和“计算机科学”的界限很难划分清楚，容易造成归类错误. 自由分类法(Folksonomy)提出了一种解决方案，即通过用户自由标注的标签来区别分类对象，被同一个标签所标记的对象归属为一类[2-3]. 这种方法具有灵活、开放的特性，但是自由标注缺乏规范制约，会引起大量的类别信息冗余，以至于造成标签管理困难等问题.

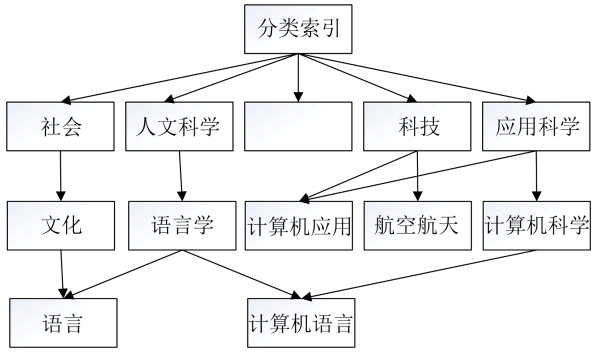

图 1　维基百科词条分类结构示意图

Fig. 1　Structural diagram of entry classification in Wikipedia

随着互联网和大数据技术的发展，面向海量数据的自动分类技术成为研究热点问题. 其中，多层次多标签分类(hierarchical multi-label classification，HMC)[4-5]是一种基于多标签层次结构的分类器，它将标签纳入层次分类系统中，每一个词条可以与类别标签层次结构中的多个路径相关，融合了分类法和标签法的优点. 设计该分类器的目的是为了利用无监管机器学习方法，实现海量数据的自动分类. “校园百科平台”参考了 HMC 的分类结构，设计了适用于校园百科的词条组织体系，如图 2 所示.

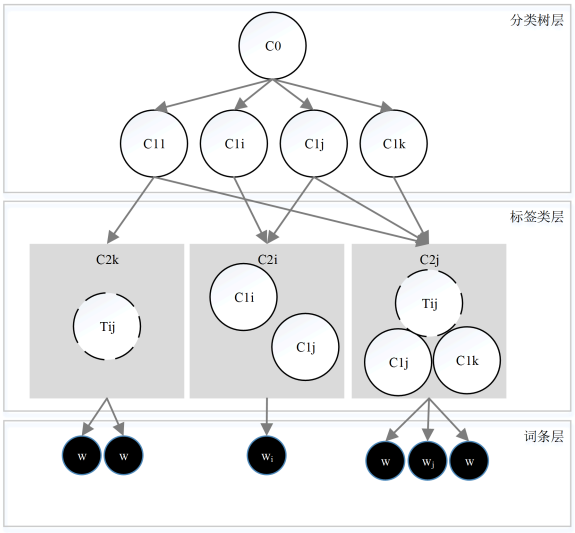

图 2　校园百科词条组织结构

Fig. 2　Structural diagram of entry classification in campus encyclopedia

校园百科词条组织结构是以分类树为基础，兼有语义标签层的分类结构. 它将多标签节点作为子类挂接在分类树上，多标签节点的所属父类是由其标签来源决定的. 由图 2 所示，“校园百科”的词条组织结构是由分类树层、标签类层、词条层组成. 除 C_0 根节点，分类树层只包含一层互斥的分类 C_1 层，避免了多层分类带来的子类别设置困难的问题. 标签类层是基于已定义的标签库的标签类，即 C_2 层. 标签库包括 C_1 层节点的扩展标签(如 T_{ij})和 C_1 层节点标签. C_2 层的标签节点标签为来自标签库的标签组合，且满足两个基本条件：①任何 C_1 层节点标签不可与自己的扩展标签属于 C_2 层同一节点；②任何 C_2 层节点如果存在 C_1 层节点标签，则该 C_2 层节点至少包含两个标签. C_2 层节点标签来源决定了 C_2 层节点的父类节点. 例如，C_{2j} 节点标签 T_{ij} 是 C_{11} 节点的扩展标签，则 C_{2j} 属于 C_{11} 节点的子类；同时它也包含 C_{1j} 和 C_{1k} 标签，因此 C_{2j} 也是 C_{1j} 和 C_{1k} 节点的子类. 词条具有完全相同标签的则被视为一类. 如图 3 所示，“乌鸫”和“灰喜鹊”具有完全相同的词条标签，则归属于同一标签节点.

在标签检索过程中，具有目标标签的所有词条将会全部返回，这里的检索返给用户面向语义的所有相关词条，打破了类别的限制. 由图 2 可知，如果搜索关键词 T_{ij} 时，则 C_{2k} 和 C_{2j} 类中的所有词条将返回给用户. 这种词条组织方式既兼容传统的基于树状图的词条分类结构，又纳入了标签类的灵活性，为建设语义网络和数据挖掘提供了高效的数据组织结构.

1.2　智能检索

校园百科平台将精确检索、标签检索词条推荐过程统一起来，突出了信息检索的个性化和智能化特征. 另外，校园百科支持全文检索，将无结构的词

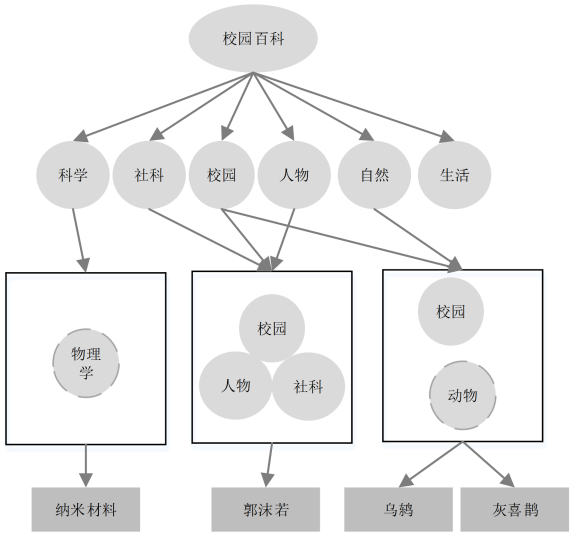

图 3 校园百科词条分类样例

Fig. 3 Sampling diagram of entry classification in campus encyclopedia

条内容转化为拓扑结构的知识网络.

1.2.1 关联检索

本文的关联检索是指由精确检索、标签检索和词条推荐组成的词条智能检索,目的是为用户提供友好的、个性化、智能化的词条检索服务和知识推荐,检索流程如图 4 所示.

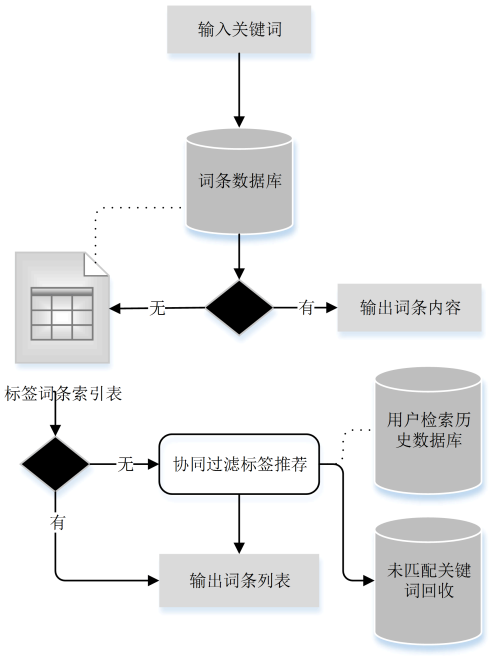

图 4 基于标签库关联检索流程

Fig. 4 Flow chart of label-based search

由图 4 可知,用户输入的检索关键词首先通过词条库进行精确检索. 如果成功,则直接返回词条内容;否则,转入基于标签的检索过程. 标签词条索引表是记录标签库的标签与词条一对多关系的列表. 标签可以视为词条的属性,用于概括词条的基本语义[6]. 若检索关键词与标签索引表中的标签的匹配成功,则返回该标签下的所有词条. 如果有多个检索关键词,则跳过精确检索,直接进行多标签匹配,结果返回所有匹配标签下的词条交集. 如果标签匹配失败,则通过协同过滤标签推荐算法为用户推荐词条;同时,将未能匹配的关键词放入词条回收站,用于词条招领.

1.2.2 全文检索

词条内容隐藏着词条之间丰富的联系,挖掘词条之间的关联,建立知识网络是校园百科建设的核心问题. 维基百科是通过超文本链接建立词条之间相互参考的网络结构. 这种网络结构的建立将非结构化的词条内容转变为词条知识网络,为关联搜索提供数据源. “校园百科”的全文搜索功能是基于 Elasticsearch 搜索引擎实现的[7-8]. 即通过对词条内容分词,并为与词条库中词条匹配的所有关键词创建 Elasticsearch 索引表,实现词条内容网络化. Elastcisearch 中文分词包括两种模式: ik_max_word 和 ik_smart. 前一种模式做最细粒度划分,如“中国科学技术大学”会拆分为“中国科学技术大学”“中国科学技术”“大学”“中国科学”“技术大学”“中国”“科学技术大学”“科学”“技术”,这种方式穷尽各种可能的拆分. 这种拆分不仅产生了过多的冗余分词,而且会占用大量计算存储资源. 后一种模式为最粗粒度拆分,那么“中国科学技术大学”则不会被拆分,被视为一个词条. 一般来说,为了最大化词条内容分词在索引时用 ik_max_word. 与其不同的是,校园百科是面向特殊领域的百科平台,涉及的大多数词条具有个性化的特点,广而普的词条并不是校园百科的主要搜索对象;另外,在知识推荐过程中,冗余的分词索引会造成检索特征提取不明确、推荐词条不准确的问题. 从用户快速准确搜索和知识推荐的角度来看,最粗粒度 ik_smart 模型适用于校园百科的词条内容分词. 首先利用 ik_smart 模型实现对词条内容分词,再利用现有的词条库与分词结果做精确匹配,过滤掉不准确或无效分词. 最后,将完全匹配的分词添加到索引表中,用于建立词条与词条内容分词的关联网络. 图 5 是利用 Nosql 图形化数据库 Neo4j,提取 100 个词条及其关联,建立的校园百科知识网络示意图.

1.3 知识推荐

校园百科平台作为校园文化交流和传播的平台,根据用户关注的热点问题进行推荐,是校园百科平台的重要功能之一. 协同过滤是目前应用最广且最为有效的一种推荐技术,它首先通过构建项目与用户行为的特征矩阵,然后计算用户间行为特征的相似性,按评分大小来进行推荐[9-11]. 由于传统的协同过滤技术面临着高维稀疏矩阵的问题,在实际应用中不容易推广. 高维稀疏矩阵为了增强特征模型的高表达性,选取了过多的特征点. 校园百科平台仅以用户检索历史为依据,设计知识推荐模型. 用户检索历史数据可以挖掘出两类特征:①词条的访问次数和时间;②词条的访问方式. 词条的访问方式主要有 3 种:关键词检索访问、内容链接访问和推荐链接访问. 词条访问次数直接反映了用户的对词条的喜好程度,而词条的检索方式间接体现了用

户对词条关注程度. 如果用户先搜索词条，再进一步通过全文检索查看它的关联词条，那么可以认为用户对词条及关联词条感兴趣；如果用户访问了推荐词条，说明推荐的知识符合用户的兴趣点. 基于上述原因，我们可以设计以词条访问次数为特征向量；并用不同的访问方式作为相关系数的方式，来调节特征向量，实现知识推荐模型的设计，其模型的具体设计如下：

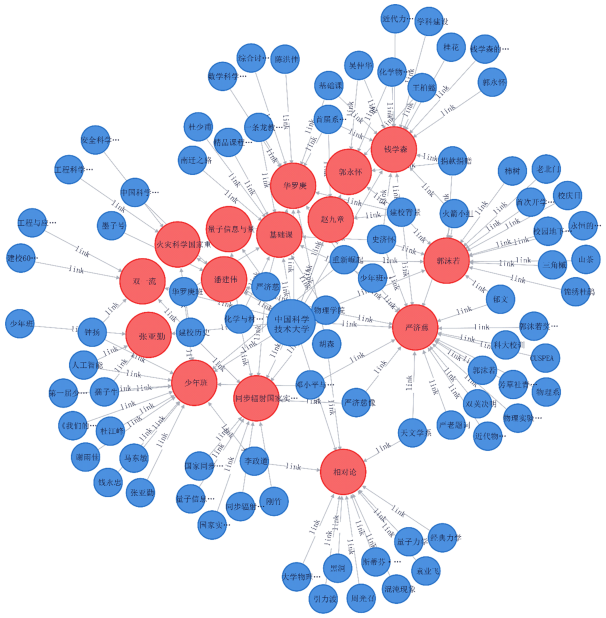

图 5　校园百科知识网络示意

Fig. 5　Sample knowledge graph of campus encyclopedia

首先，为每个用户建立基于词条库的特征向量 $(x_1,\cdots,x_i,\cdots,x_n)$，其中 n 是词条库的总数，x_i 为第 i 个词条历史检索次数的加权累加和. 这里，权值 ω 是访问词条相关系数，且 $\omega \geqslant 1$，被用于将词条的访问方式特征纳入模型. 在用户一轮词条访问过后，特征向量 $(x_1,\cdots,x_i,\cdots,x_n)$ 中任何被检索的关键词 i，将以公式

$$x_i = x_i + \omega \tag{1}$$

进行累加. 如果任何一对访问词条 (i,j) 组合存在索引关系，则分别累加 x_i 和 x_j 大于 1 的 ω；如果不存在索引关系，ω 取值为 1. 循环上述过程，直到所有被检索的关键词对完成上述操作. 另外，如果特征向量中的词条 i 是通过推荐方式访问的，那么，令

$$x_i = \omega x_i \tag{2}$$

式中，ω 取大于 1 的值.

用户的检索特征向量设置完成后，将进行用户检索相似度计算. 用户 u 和用户 v 的检索特征向量的相似度计算公式为

$$\mathrm{Sim}(u,v) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{x_i^{\,2}} \times \sqrt{y_i^{\,2}}} \tag{3}$$

式中，用户 u 和 v 分别为基于检索词条库的 n 维特征向量 $(x_1,\cdots,x_i,\cdots,x_n)$ 和 $(y_1,\cdots,y_i,\cdots,y_n)$. 通过计算目标用户与其他用户的搜索特征向量的相似度，获取排名最高的用户搜索特征，并将其近期搜索频率最高且目标用户仍未检索的词条作为推荐结果推荐给目标用户. 随着校园百科使用逐渐普及，知识推荐也会更加精确化和个性化.

2　校园百科架构

校园百科平台既是一个多人协作的信息编辑系统，也是一个基于知识网络的信息发布和检索系统. 中国科学技术大学校园百科的架构采用了 4 层结构：数据层、业务层、RESTful API 和展示层，如图 6 所示.

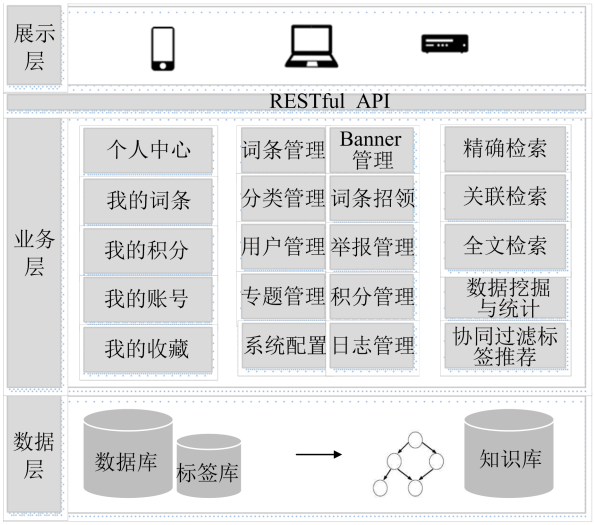

图 6　中国科学技术大学校园百科系统架构

Fig. 6　Structure diagram of campus encyclopedia of USTC

由图 6 可知，数据层用于存储词条、标签和索引数据，并支持 NoSQL 数据库存储知识图谱. 业务层可以分为面向个人用户的功能、面向系统管理的功能、面向公共服务功能. 百科平台允许个人用户编写、维护、收藏词条，管理个人账户与积分；允许管理员管理词条和类别标签，设计专题，发布海报和招领词条以及用户和权限管理、积分和举报管理；

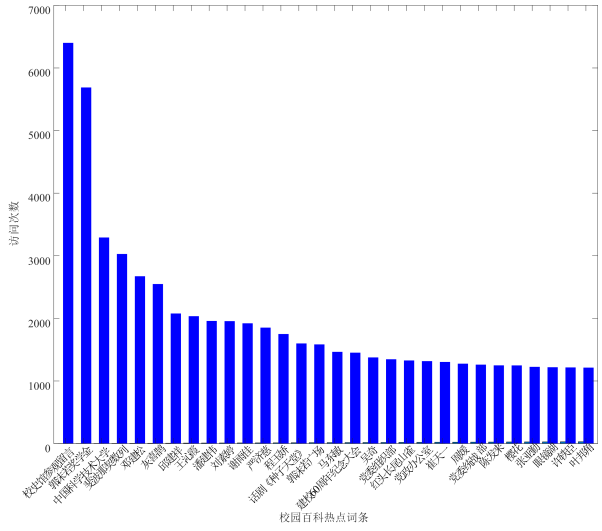

图 7　校园百科热点词条统计图

Fig. 7　Statistics of popular search entries in campus encyclopedia

为了提高系统和数据的可拓展性、可移植性，校园百科采用了 RESTful API 为用户提供灵活的数据访问接口.展示层可实现面向 PC 端的用户接口、面向移动端的用户接口以及面向其他设备的服务访问接口.校园百科作为校园文化知识库，已为我校的智能虚拟助手提供了数据支持，方便了我校师生随时随处检索词条、学习校园文化.

目前，校园百科在我校还处于推广的起步阶段，积累了 879 例词条，实现编写 1435 次，共 277 人参与.此后，我们将持续采用积分兑现和院系合作的方式，推动校园百科词条的累积，推广校园百科平台的使用.图 7 展示了前 30 个最受关注的热点词条，特别是“校史管参观留言”和“郭沫若奖学金”访问次数均超过了 5500 次.

3 结论

随着互联网进入 Web2.0 时代，基于知识推荐的校园百科是智慧校园建设的组成部分，也是新一代校园文化交流和分享的平台典型范例.它不仅支持学校师生协作编写各类校园文化内容，实现校园文化知识积累和共享；而且它还支持多渠道访问接口和应用拓展，满足了师生随时随地访问各类校园知识和服务的需求.校园百科平台具有创新性、实用性和推广性的特征，值得在智慧校园建设中继续探讨和推广.

参考文献(References)

[1] 王静. 中文维基百科类别推荐研究[D]. 武汉：华中师范大学，2016.

[2] DYE J. Folksonomy：A Game of High-tech (and High-staked) Tag [J]. E Content，2006，29(3)：38-43.

[3] VOSS J. Collaborative thesaurus tagging the Wikipediaway[EB/OL]. [2006-04-27]. http://arxiv.org/abs/cs.IR/0604036.

[4] SILLA C N，FREITAS A A. A survey of hierarchical classification across different application domains [J]. Data Mining & Knowledge Discovery，2011，22(1-2)：31-72.

[5] READ J，PFAHRINGER B，HOLMES G，et al. Classifier chains for multi-label classification [J]. Machine Learning，2011，85(3)：333.

[6] 孙红莺，次仁拉珍，叶鹰. 基于标签的数字图书馆个性化信息检索[J]. 杭州师范学院学报：自然科学版，2008，7(5)：387-391.

[7] HATCHER E，GOSPODNETIC O. Lucene in Action [M]. USA：Manning Publications Co.，2004：50-350.

[8] KONONENKO O，BAYSAL O，HOLMES R，et al. Mining modern repositories with elasticsearch [C]// ACM，2014：328-331.

[9] HERLOCKER J L，KONSTAN J A，TERVEEN L G，et al. Evaluating collaborative filtering recommender systems [J]. ACM Transactions on Information Systems (TOIS)，2004，22(1)：5-53.

[10] 周万珍，曹迪，许云峰等. 推荐系统研究综述[J]. 河北科技大学学报，2020，41(1)：76-87.

[11] 肖运文. 基于 ElasticSearch 的教育资源推荐系统设计与实现[D]. 北京：北京工业大学，2016.