DOI:10.13718/j.cnki.xsxb.2017.07.013

基于数据挖掘算法的电子图书馆智能推荐技术研究①

张 容, 张 勇

成都医学院 图书馆, 成都 610500; 四川职业技术学院 计算机科学系, 四川 遂宁 629000

摘要: 针对电子图书馆的智能推荐服务, 提出了一种基于数据挖掘算法的新方法. 此方法采用中图分类号索引树计算读者的兴趣倾向程度, 采用改进的 K-means 聚类方法实现兴趣相近的读者聚类, 采用改进的 Apriori 算法实现关联规则挖掘并形成智能推荐建议. 此方法在电子图书馆的实际应用中, 为读者提供了预期的推荐服务, 3 个等级的满意度达到了 92.8%.

关 键 词: 数据挖掘; 电子图书馆; 智能推荐; 聚类分析

中图分类号: TP391 **文献标志码**: A **文章编号**: 1000-5471(2017)07-0081-05

信息技术的飞速发展, 推动了图书馆的信息化和电子化. 电子图书馆的出现, 不仅大大增加了馆藏资源的总体容量, 也便利了读者对图书馆资源的访问和下载[1].

电子图书馆在发展过程中, 也遇到了新的挑战. 最受广大读者关注的问题之一就是, 如何从电子图书馆的海量电子图书资源中, 快速、准确地搜索到自己所喜爱的书目. 这一问题的出现, 为电子图书馆提出了一个崭新的服务需求——为用户提供智能推荐服务[2-3].

数据挖掘技术是信息搜集、数据处理、深度语义关联检索中的常用技术. 如果将此技术引入电子图书馆建设之中, 根据用户的历史下载记录, 结合数据挖掘算法中设定的规则, 就使得在海量电子图书资源中快速检索到用户的需求书目成为可能[4]. 为此, 本文以电子图书馆建设为背景, 依托数据挖掘算法构造一种智能推荐技术, 以适应电子图书馆发展的需要.

1 电子图书馆的智能推荐服务设计

在电子图书馆中, 要实现图书资源满足读者需求的智能推荐, 是一种典型的机器学习、智能识别、数据挖掘的过程[5]. 因此, 根据读者提供的明确查询信息或者读者过往的下载数据, 对读者兴趣倾向进行判断, 再根据关联规则进行数据挖掘, 最终形成推荐建议是智能推荐服务的基本设计思路.

1.1 读者倾向程度判定

根据读者要下载的图书信息或过往下载的图书信息, 可以形成对读者兴趣倾向的判断.

电子图书馆中的图书具有很多信息标注, 如书名、ID 号、ISBN 等等. 其中, 中图分类号是判断 1 本图书类别属性最简单明了的信息. 随着中图分类号的层层深入, 可以构造出 1 本图书索引树, 如图 1 所示.

如果 2 名读者下载的图书是同一本书, 或者 2 本图书具有相同的中图分类号(书名不同), 则可以根据

① 收稿日期: 2016-12-28
基金项目: 四川省学术成果分析与应用研究中心课题项目(SCAA15B12).
作者简介: 张 容(1979-), 女, 四川南充人, 硕士, 讲师, 主要从事英语教育与图书馆信息服务研究.

下载时间的差异来评价读者之间的兴趣倾向是否相近，据此建立的数学模型如公式(1)所示.

$$Sim(A,B)=\frac{\frac{1}{2}(t_{1A}+t_{2B})}{\frac{1}{p+q}\left(\sum_{i=1}^{p}t_{1i}+\sum_{i=1}^{q}t_{2i}\right)} \quad (1)$$

这里，A、B表示中图分类号相同的2本图书；t_{1A}表示了读者R_1下载图书A的时间；t_{2B}表示了读者R_2下载图书B的时间；t_{1j}表示了读者R_1下载图书j的时间；t_{2j}表示了读者R_2下载图书j的时间；p,q表示了下载次数.

如果2名读者下载的图书不是同一本书，并且中图分类号也不相同，则需要根据图书所对应的中图分类号在索引树中的深度来判断读者之间的兴趣倾向是否相近，据此建立的数学模型如公式(2)所示.

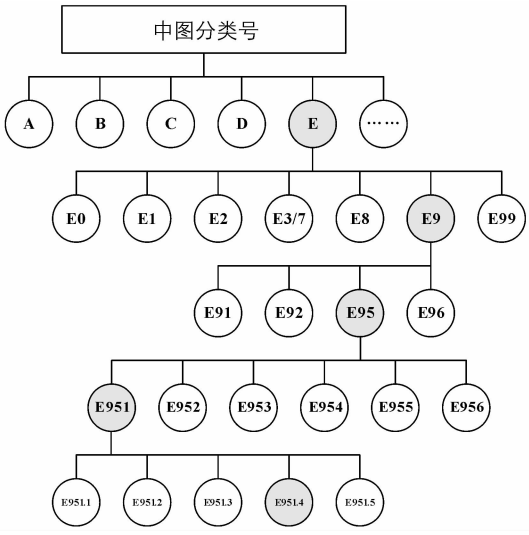

图1 基于中图分类号的图书索引树

$$Sim(A,B)=\frac{D(Z(A,B))}{(D(A)+1)+(D(B)+1)} \quad (2)$$

这里，$D(A)$表示图书A在中图分类号索引树中所处的深度；$D(B)$表示图书B在中图分类号索引树中所处的深度；$Z(A,B)$表示图书A和B在索引树中最接近的节点；$D(Z(A,B))$表示$Z(A,B)$在索引树中所处的深度.

统计出2个读者全部的具有相似性的图书后，可以计算出其兴趣倾向的相似程度.

$$C(R_1R_2)=\frac{1}{1+\sqrt{\sum_{i=1}^{n}Sim_i^2}} \quad (3)$$

1.2 相似倾向读者聚类

如果能将读书兴趣倾向相似的读者都归并到相同类别中，那么同一类读者群体的兴趣倾向是为类内读者提供推荐的很好判据. 所以，数据挖掘领域中的聚类分析对于电子图书馆的智能推荐服务具有重要意义[6].

K-Means算法是聚类分析中的经典算法，但其聚类中心随机设置，导致聚类分析的结果不一致等问题.

据此，本文将图割理论引入到K-Means算法中，提出一种适用于读者兴趣聚类的新算法，算法的流程如下：

第一步，构建读者集合$\{R_i\}$，$\{R_i\}$中的元素成为图的各个顶点，$\{R_i\}$中任意2个元素之间的连线成为图的各个边，这些边构成一个集合$\{E_j\}$. 至此，可以构建一个无向连通图$G=(\{R_i\},\{E_j\})$.

第二步，根据公式(1),(2),(3)计算$\{R_i\}$中任意2个读者的兴趣相似程度，并以此设置$\{E_j\}$的各个边. 这样，G变成了一个带权无向连通图.

第三步，根据各个边的权重进行由大到小的排列. 选取最大权重的边连接2个对应顶点，形成第一个连通图分量. 再选择次大权重的边所对应的2个顶点，形成第二个连通图分量. 依次类推，顺次执行. 当新扩展的连通分支和已有的连通图分量无交叉时，就按既定形状进行扩展. 当新扩展的连通分支和已有的连通图分量形成交叉时，此扩展无效. 当新扩展的连通分支和已有的连通图分量形成闭合回路时，就形成了一个聚类，并将其纳入聚类集合$\{\vartheta_i\}$中.

第四步，计算每个聚类ϑ_i的中心o_i，其公式为

$$o_i=\sum\frac{E_i}{N_{R_i}} \quad (4)$$

这里，N_{R_i} 表示 ϑ_i 中读者的数量.

第五步，根据公式(1)，(2)，(3) 重新计算读者到各个聚类中心 o_i 的距离，并根据最大相似性原则再次执行聚类划分.

第六步，重复上述过程，直至聚类中心不再变化为止.

1.3　改进 Apriori 关联规则挖掘

经过聚类分析后，有相似兴趣倾向的读者被归并到一个类别中，这个读者类别所关联的图书资源也就成为提供智能推荐的初始依据. 读者之间的关联、图书之间的关联，还有哪些深层次的隐含规则，则需要进一步利用数据挖掘算法去实现.

Apriori 算法是关联规则挖掘的常用方法，但其形成的关联规则很多是冗余的，并且需要执行的扫描次数也比较多[7]. 为此，本文在传统 Apriori 算法上进行改进，改进算法的流程如下：

第一步，对数据集合 DataBase 执行扫描处理，扫描的目的是获得一个不含有任何重复元素的候选条目集合. 在此步骤中，与传统 Apriori 算法不同的是，设置一个累加计数，这个累加计数将作为候选频繁集合的支持度，用于后续的剪枝处理，进而得到频繁集合 1 -项集 L_1.

第二步，对频繁项集 1 -项集合 L_1 执行自链接处理操作，形成候选频繁 2 -项集合 C_2. 遍历候选频繁 2 -项集合 C_2 中的所有条目项，根据最小支持度执行剪枝处理，即支持度大于 2 的条目项可以保留，支持度小于 2 的条目项则不保留，最终获得频繁项集 2 -项集合 L_2，同时建立一个对应于 L_2 的地址映射列表 A_2，这个对应关系如图 2 所示.

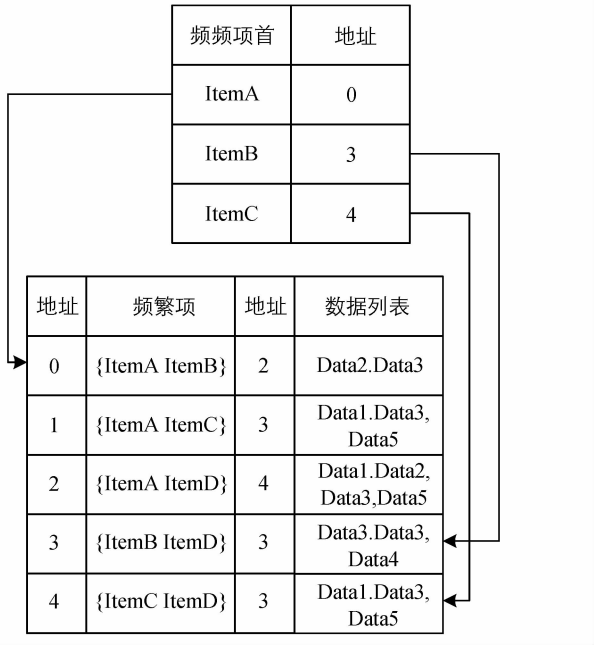

图 2　L_2 对应的地址列表 A_2

第三步，当自链接操作 $L_{(k-1)}$ 用于生成频繁集 C_k 时，通过对其对应的地址列表 $A_{(k-1)}$ 执行搜索来完成. 根据地址的搜索，如果发现 C_k 中的某个选项并非属于 $L_{(k-1)}$ 的子集，则直接将这项删去. 如果满足搜索条件的，则加入新的自链接操作集 L_k 中，同时映射生成对应的 A_k.

第四步，反复执行第二步和第三步，直到确认 C_k 为空集.

2　电子图书馆的推荐服务实现及效果评价

依据本文所提出的聚类分析、关联规则挖掘等数据挖掘算法，结合数据预处理等技术，本文构建了一个面向高校的电子图书馆系统，并于 2016 年在我校图书馆试运行.

2.1　电子图书馆的总体设计

在 B/S 通信模式下，电子图书馆分别从客户端和服务器端进行设计.

客户端包含了读者帐号注册模块、读者帐号维护模块、电子图书浏览模块、电子图书检索模块、电子图书下载模块、电子图书预订模块.

服务器端包含了图书资源管理模块、读者信息管理模块、智能推荐管理模块、图书预订管理模块等. 下面将主要介绍智能推荐模块的实现效果.

2.2　推荐服务的实现结果

聚类分析是电子图书馆智能推荐服务的基础性工作，在本文的智能推荐模块中，结合 1.1 节和 1.2 节的工作实现了聚类分析. 如表 1 所示，就是部分读者的聚类分析结果.

表1 部分读者的聚类分析结果

读者编号	读者姓名	聚类结果	兴趣备注
89020305	王 洋	Cluster1	爱好历史
88090122	杜 鹏	Cluster1	爱好历史
87010708	李向阳	Cluster2	爱好文学
89040315	张达斌	Cluster1	爱好历史
90050631	陆 丰	Cluster5	爱好军事
91050318	薛晓丽	Cluster7	爱好政治
82020109	王 强	Cluster6	爱好文艺
83080217	秦关山	Cluster4	爱好自然
90020305	张悦心	Cluster5	爱好军事
……	……	……	……

从表1的结果可以看出，根据读者的过往下载信息，这9名读者经过本文的聚类分析算法，被划归到各自的类别中. 并且经过聚类，这些读者的兴趣爱好也清晰地显示出来. 例如，王洋就是一个爱好历史类题材图书的读者. 经过聚类分析工作，为读者提供的推荐服务就更有针对性了.

根据1.3节的改进Apriori算法，进一步执行关联规则挖掘，得到的智能推荐建议如图3所示.

从图3中的结果可以看出，智能推荐服务根据读者“王洋”的历史下载信息进行聚类分析和关联规则挖掘后，为其推荐了6部历史题材的电子图书. 这是符合读者“王洋”的兴趣爱好的，从而证明了本文为电子图书馆构建的智能推荐服务达到了预期效果.

图3 智能推荐结果

2.3 智能推荐的满意度评价

为了进一步检验电子图书馆所提供的智能推荐是否让读者满意，本文在智能推荐服务之后进行了满意度评价.

读者可以选择的满意度评价有5个等级，即非常满意、满意、基本满意、不满意、很不满意[8-9]. 读者可以根据自己的需要，结合电子图书馆推荐的书目来进行打分.

在2015年11月—2016年1月之间，共有221位读者参与了智能推荐的满意度评价工作. 评价结果如表2所示.

表2 智能推荐的满意度评价结果

等 级	读者人数	占比/%
非常满意	58	26.2
很满意	106	48.0
满意	41	18.6
不满意	12	5.4
很不满意	4	1.8

从表2的结果可以看出，对于本文电子图书馆的智能推荐服务，满意、很满意、非常满意3个等级的读者之和达到了205人，占比达到了92.8%，这说明此电子图书馆所提供的智能推荐服务，还是让大多数读者满意的，这也证实了本文提出的基于数据挖掘算法的电子图书馆智能推荐技术的有效性.

3 结 语

电子图书馆已经逐渐成为读者喜爱的全新服务形式，其大容量、快捷性、渐变性都受到读者的青睐.

当前，继续解决的问题就是如何让读者从海量电子图书资源中快速地检索到自己想要的资源.

针对这一问题，本文对电子图书馆的智能推荐技术展开研究. 首先，根据读者下载图书的历史数据，使用中图分类号索引树进行读者兴趣倾向程度计算. 其次，对经典的 K-means 聚类方法改进，以实现兴趣倾向近似的读者信息聚类. 最后，对 Apriori 算法进行改进，实现更高效的关联规则挖掘，为读者提供更加理想的智能推荐服务.

依托此智能推荐技术，构建了电子图书馆. 此电子图书馆在实际使用中的效果，证实了本文方法的有效性.

参考文献：

[1] 马晓亭. 复杂云计算环境下基于客户感知价值的数字图书馆服务效能评估 [J]. 图书馆理论与实践，2014(3)：84－86.

[2] LOPS P，GEMMIS M D，SEMERARO G，et al. Content-based and Collaborative Techniques for Tag Recommendation：an Empirical Evaluation [J]. Journal of Intelligent Information Systems，2013，40(1)：41－61.

[3] 张建娥. 基于 TFIDF 和词语关联度的中文关键词提取方法 [J]. 情报科学，2012，30(10)：1542－1555.

[4] 郝小花，邓小昭. 基于数据挖掘的可视化数字图书馆用户社区聚类与特征分析 [J]. 情报科学，2008，26(3)：396－399.

[5] TINTAREV N，MASTHOFF J. Evaluating the Effectiveness of Explanations for Recommender Systems [J]. User Modeling and User-Adapted Interaction，2012，22(4/5)：399－439.

[6] 冯英华，刘　磊. 基于需求的高校图书馆 2.0 个性化信息服务模式研究 [J]. 中国图书馆学报，2012，38(2)：50－61.

[7] KANI-ZABIHI E，GHINEA G，CHEN S Y. Digital Libraries：What do Users Want? [J]. Online Information Review，2006，30(4)：395－412.

[8] 李章平，陈玉成，罗林坤. 图书馆读者满意度测评中的 3 种定性因子赋权方法 [J]. 西南师范大学学报(自然科学版)，2010，35(2)：242－246.

[9] LI Y，WONG I S M，CHAN L P Y. Mylibrary Calendar：A Web 2.0 Communication Platform [J]. Electronic Library，2010，28(3)：374－385.

On Intelligent Recommendation Technology of Electronic Library Based on Data Mining Algorithm

ZHANG　Rong[1]，　ZHANG　Yong[2]

1. *Library of Chengdu Medical College，Chengdu 610500，China*；

2. *Department of Computer Science，Sichuan Vocational and Technical College，Suining Sichuan 629000，China*

Abstract：A new method based on data mining algorithm has been proposed for the intelligent recommendation service in the electronic library. By this means，tree computing reader's interest tendency has been indexed，by improved k-means clustering method，to realize the clustering of similar interest readers by improved Apriori algorithm realized the association rules mining and intelligent recommendation. This method in the practical application of the electronic library，to provide readers with the expected service，three grades of satisfaction reached 92.8%.

Key words：data mining；electronic library；intelligent recommendation；cluster analysis

责任编辑　夏　娟