数字图书馆

基于多兴趣特征分析的图书馆个性化图书推荐方法

马 健 杜泽宇 李树青
(南京财经大学信息工程学院 南京 210046)

【摘要】应用渐进遗忘策略和滑动窗口相结合的更新算法等,建立读者的兴趣词库和索引库,进而建立读者的多兴趣特征库。分别计算读者兴趣特征的特征词库以及索引库与书籍的相似度,将这两种方法计算出的相似度进行线性叠加,建立具有可操作性和扩展性的混合推荐算法,从而实现图书馆书籍的个性化推荐方法。该方法综合利用《中图法》中书籍所属的索引类别,能有效解决数据稀疏问题。最后对相关实验内容和结果进行详细说明。

【关键词】个性化推荐 渐进遗忘策略 兴趣特征 混合推荐
【分类号】TP391

Personalized Book Recommendation Algorithm Based on Multi-interest Analysis in Library

Ma Jian Du Zeyu Li Shuqing
(College of Information Engineering, Nanjing University of Finance & Economics, Nanjing 210046, China)

【Abstract】This paper firstly constructs the multi-interest feature library from readers' interest lexicon and index with update algorithms combining gradual forgetting strategy and sliding window, then calculates the similarity measures of readers' interest lexicon and index with books, and adds the two similarity with linear superposition to propose an operable and extensible hybrid recommendation algorithm. This algorithm synthetically uses the index types of books in Chinese Library Classification, and effectively solves the problem of data sparseness. Finally, the paper achieves a personalized recommendation system of the library books, and correlative experimental results are introduced in details.

【Keywords】Personalized recommendation Gradual forgetting strategy Interest feature Hybrid recommendation

1 引 言

图书馆是读者获取图书资源的重要途径,网络技术的发展给图书馆信息服务带来了新的挑战。传统的信息服务模式下,所有的读者面对同一平台,需要读者主动提交查询请求来搜索自己所需要的信息,但随着高校图书馆信息量的膨胀,读者在传统的信息服务模式下很难获得符合其兴趣特征的信息。个性化推荐服务则是根据不同读者的兴趣特征,主动帮助读者从大量的信息中寻找出可能感兴趣的信息,并及时推荐给读者。目前个性化推荐服务被应用在很多领域,如电子商务、Web 信息检索等[1],其中应用较成熟的是基于协同过滤的推荐技术。但在其应用过程中仍存在很多问题,如读者评价矩阵的稀疏性、算法的可扩展性等[2]。更重要的是图书馆的读者都

收稿日期:2012-04-27
收修改稿日期:2012-06-10

有特定专业背景,读者兴趣特征有其不同于电子商务和 Web 信息检索等环境的特殊性,单纯使用上述模型难以得到满足读者需求的个性化推荐系统。

本文结合图书馆书籍借阅的具体特征,在建立读者特征库和计算文本相似度等的基础上,提出了适用于高校图书馆图书个性化推荐的模型,并对相关数据进行实验分析,从而达到为读者提供符合其兴趣特征的图书资源的目的。

2 文献回顾

1991 年美国启动的数字图书馆计划是图书馆个性化信息服务的先驱,接着日本、英国等国家也耗巨资开展数字图书馆的研究[3]。1999 年美国图书馆与信息技术联合会(Library Information Technology Association, LITA)把个性化定制服务列为数字图书馆技术发展 7 大趋势之首[4,5]。图书馆中书籍的个性化推荐属于个性化服务的一种,个性化推荐服务是一种深层次的、主动性和个性化较强的服务方式[6]。随着网络的普及,数字化的信息越来越多,信息技术尤其是人工智能技术的进步,图书馆个性化推荐的研究受到国内外越来越多学者的关注[7,8]。根据推荐算法的不同,图书馆书籍的个性化推荐方法可以分为三类,即协同过滤推荐方法、基于内容的推荐方法以及混合推荐方法[9,10]。

协同过滤推荐方法的核心思想是利用读者对书籍的历史借阅信息,生成和目标读者兴趣接近的邻居集,再根据生成的邻居集中的读者对目标读者生成推荐。协同过滤推荐系统最大的优点是对推荐对象没有特殊要求,能处理音乐、电影等难以进行结构化表示的对象。缺点主要是稀疏性问题,即在系统使用初期,由于系统资源还未获得足够多的读者对于书籍的评价,系统很难利用这些评价来发现相似的读者[11]。有学者应用改进用户评价矩阵的协同过滤推荐[12],但没有考虑书籍之间的相关性。还有学者基于协同过滤的思想提出应用关联规则寻找最大频繁模式的个性化推荐[13,14],但这种方法对读者个人的借阅信息考虑较少,实际中难以满足读者对书籍的个性化需求。

基于内容推荐方法是通过读者以往借阅过的书籍的相关信息与现有书籍之间的匹配程度进行过滤推荐。基于内容推荐的优点主要是可以处理新用户和新项目问题,即冷启动问题。但是也面临很多新问题,如推荐结果比较单一,只能推荐出用户特定的兴趣内容,无法挖掘用户的潜在兴趣。有学者提出了结合内容相似度与聚类的推荐技术[15],还有学者提出了基于内容的蚁群聚类协同过滤推荐算法[16],但这些算法对读者的个人借阅信息考虑不足。

鉴于协同过滤推荐和基于内容推荐两种方法均有各自的优缺点,许多学者从不同角度提出了基于这两种方法的混合模型。例如,有的学者先利用回归算法提取书籍的内容属性,再对读者进行基于内容的协同过滤,最后同构化整合二者结论,并进行加权求和与有序化,从而提出了一种基于内容和协同过滤同构化整合的推荐系统模型[17]。还有学者应用模糊聚类技术,结合读者项目矩阵和类似的项目多层次的关联,综合提出基于内容和协同过滤的推荐算法[18]。但这些方法本身考虑的是电子商务的应用,难以直接移植到图书馆书籍的个性化推荐中。本文提出的综合考虑读者个人特征以及借阅记录和书籍相关性,建立读者特征库的模型也是一种混合推荐模型。

3 基本思路

本文根据图书馆书籍借阅的具体特点,综合考虑多种合理因素建立模型。读者的个性化需求与读者借阅书籍的名称、借阅书籍的索引号、书籍借阅的时间、读者的专业、年级、读者当前的课程安排以及当前馆藏书籍的实际情况等因素都有关,本文试图全面考虑多种影响因素,建立真实全面、具有可操作性和可扩展性的数学模型。

3.1 读者兴趣模型的初始化构造

读者兴趣的初始化模型根据读者的历史借阅记录、读者的专业以及当前的课程安排等因素确定。从读者的历史借阅记录中借阅书籍的名称、读者所在专业的核心课程、读者当前学期的主要课程等提取特征词,根据借阅记录中每本书籍的借阅序列以及借阅时间等信息确定每个特征词的权重,每个特征词表示一个兴趣。一个兴趣可以被多个读者共同拥有,一个读者也可以拥有多个兴趣,即兴趣与读者之间是多对多的关系。系统从读者借阅书籍的索引号中提取索引分类,根据每个分类的借阅情况确定每个索引分类的权重。

初始的读者兴趣模型 I^0 可以表示为一个二元组:

$$I^0 = (W, P)$$

其中，W 表示读者特征词库，P 表示读者的索引库。

读者的特征词库 W 也是一个二元组，由 W_t，W_f 两部分组成：

$$W = (W_t, W_f)$$

其中，W_t 表示根据读者的历史借阅记录中借阅书籍的名称、读者所在专业的核心课程以及读者当前学期的主要课程确定的特征词，W_f 为该特征词对应的权重。

读者的索引库 P 也是一个二元组，由 P_t，P_f 两部分组成：

$$P = (P_t, P_f)$$

其中，P_t 表示读者历史借阅的书籍所隶属的各级索引分类号，P_f 表示该索引分类对应的权重。

读者的特征词库以及索引库中的权重的计算都要用到读者的借阅书籍索引特征 F，它由 F_t，F_i，F_d 三部分组成：

$$F = (F_t, F_i, F_d)$$

其中，F_t 表示借阅书籍对应的索引号；F_i 表示该借阅书籍在该读者全部借阅书籍中的序列；F_d 表示该借阅书籍对应的借阅时间。

3.2　特征词库索引库的建立以及权值的计算

（1）特征词库中特征词的提取

首先考虑读者借阅的书籍名称，每本书的书名都可以分解为几个词组，而每个词组就是潜在的特征词，本文中特征词的提取采用开源的基于 Java 语言开发的轻量级的中文分词工具包 IK Analyzer，它特有正向迭代最细粒度切分算法[19]，该算法支持细粒度和最大词长两种切分模式，引入简单搜索表达式，采用歧义分析算法优化查询关键字的搜索排列组合，能极大地提高检索的命中率。

对于每位读者，他所在的专业学科本身有核心课程，例如计算机科学与技术专业[20]，这些课程本身也可以使用上述正向迭代最细粒度切分算法提取特征词。每个学期读者有当前学期的主要课程，这些课程本身也可以使用上述正向迭代最细粒度切分算法提取特征词。本文分别使用上述三种途径获取读者的特征词，选取几个专业的数据，经实验分析发现从读者借阅的书籍、专业学科、当前课程三种途径获取的特征词有一定的重合和正交相关性，即一个特征词可能从这三种途径中任一种都可能得到。本文经过分析认为，读者历史借阅记录中提取出的特征词表明读者对该特征词感兴趣，而从专业核心课程和读者当前主要课程中提取的特征词从一定程度上反映了该特征词的重要性。对于一个特征词，能获得的途径越多表明它的先天重要性越大，本文确定特征词 j 的先天重要系数 $r_j \in \{1, 2, 3\}$，分别表示特征词在这三种途径中出现 1、2、3 次。

（2）特征词库中特征词权重的计算

特征词库中特征词的权重与特征词的先天重要系数、读者借阅书籍的借阅序列、借阅时间等因素有关。对于借阅序列，采用渐进遗忘和滑动窗口相结合的更新算法，计算借阅序列对特征词权值的影响；对于借阅时间，利用该本书籍的借阅时间和读者借阅的所有书籍的平均借阅时间的比值确定借阅时间对特征词权重的影响。

①渐进遗忘和滑动窗口相结合的更新算法

在实际的系统中读者兴趣序列庞大，因此必须限定读者兴趣的总数量，当新加入兴趣和以前兴趣的数量之和大于规定的总数量时，就必须考虑将部分兴趣移出。可以采用滑动窗口方法进行处理，规定窗口的大小为 L，当有多于 L 个兴趣出现时，单纯的滑动窗口是按照到来的先后顺序，将最初到达的兴趣移出，渐进遗忘和滑动窗口相结合的方法是将对读者重要性最小的一个兴趣移出窗口。兴趣对读者新信息的推送算法的重要性即兴趣的权重，可以用遗忘函数计算，由于人对事物的遗忘是一个渐进的过程，因此遗忘函数是一个连续递减的函数。

定义渐进遗忘函数 $c = f(t)$ 表示兴趣的权重随时间的变化，此函数对每个观测到的兴趣特征根据其出现的时间次序生成权重。采用的线性函数如下[21]：

$$c_i = -\frac{2k}{n-1}(i-1) + 1 + k \qquad (1)$$

其中，n 为借阅序列的长度，假定借阅时间在同一天的书籍属于同一个序列号。i 为计数值，$i \in \{1, 2, 3, \cdots, n\}$ 按照从最近选择的特征到第一次选择的特征的顺序，$k \in [0, 1]$ 表示遗忘的快慢。当 $k = 0$ 时，表示没有遗忘；当 $k = 1$ 时，相对于滑动窗口，表示完全遗忘。对每一个特征 j，每次有新的特征值出现时，重新计算所有特征的重要性，其中重要性最小的被遗忘，从而计算借阅序列对特征词的权值的影响 FI_j[21]：

$$FI_j = \sum_{k=1}^{m} c_k a_k^j \qquad (2)$$

其中，FI_j 表示计算借阅序列对特征词 j 的权值的影响，i 为计数值，m 为读者的借阅书籍的行为数，$a_k^j \in \{0, 1\}$ 表示某

次计算时此特征是否出现，出现此特征，则 $a_k^j=1$，否则 $a_k^j=0$。c_k 为遗忘函数计算的读者某次借阅的书籍对应的借阅序列的权重。

②书籍借阅时间对特征词的权值的影响

读者借阅一本书籍都要保留一段时间，借阅时间的长短反映读者对该本书籍的喜好程度，本文认为读者对某本书籍的借阅时间越长表示对该本书籍喜好程度越高，从而对该本书籍所包含的读者的特征词的兴趣越大。所以对于每本借阅书籍 i，考虑该书籍的借阅时间和读者借阅的所有书籍的平均借阅时间的比值，从而计算借阅时间对特征词的权值的影响 FD_i：

$$FD_i=\frac{F_{d_i}}{\frac{1}{m}\sum_{i=1}^{m}F_{d_i}} \tag{3}$$

其中，FD_i 表示借阅时间对特征词 i 的影响，m 为读者的借阅书籍的行为数，F_{d_i} 表示书籍 i 的借阅时间，分母表示读者所有借阅书籍的平均借阅时间。

③特征词库中特征词的权重的计算

借阅序列、借阅时间对特征词的权重的影响确定以后，每个特征词可能出现在多本借阅书籍中，考虑对于每一个特征词 j，其权重是所有包含该特征词的书籍名称对应的序列权重与时间权重之积的和，每个特征词的权重表达式如下：

$$W_{f_j}=\sum_{i=1}^{m}(c_i\times FD_i\times a_i^j\times r_j) \tag{4}$$

其中，W_{f_j} 表示特征库中特征词 j 对应的权重，c_i 表示该本书籍对应的序列权重，FD_i 表示该本书籍对应的时间权重。$a_i^j\in\{0,1\}$ 表示某次计算时此特征是否出现，如出现此特征，则 $a_i^j=1$，否则 $a_i^j=0$。$r_j\in\{1,2,3\}$ 表示该特征词的先天重要系数，分别表示特征词在三种途径中出现的次数是 1、2、3 次。W_{f_j} 表示所有包含该特征词的书籍名称对应的序列权重与时间权重之积的和。根据该方法即可求出每位读者特征词库中每个特征词的权重。

(3)基于《中图法》的索引库中索引分类的提取

《中图法》是我国第一部集中全国图书馆和信息部门的力量共同编制的一部综合性大型文献分类法。《中国分类主题词表》收录了 5 万余条类目、21 万条主题词及主题标题，包括哲学、社会科学和自然科学各个领域的学科和主题概念[22]。本文将读者每本借阅书籍所隶属的从粗到细的各级索引分类。

显然每位读者的索引库是由多棵树组成的，从上往下分类越来越细，所有借阅的书籍都是每棵树的叶子节点。

(4)索引库中索引分类的权重的确定

每个借阅记录的索引都隶属于一个从粗到细的索引分类，对于每个索引分类 j，采用向上洪泛的计算方法计算每个索引分类的权重 P_{f_j}：

$$P_{f_j}=\sum_{i=1}^{m}\left(\frac{h}{2^{l_p}}\times c_i\times FD_i\times k_i^j\times r_j\right) \tag{5}$$

其中，P_{f_j} 表示索引分类 j 的对应的权重，i 为计数值，m 为读者的行为数，c_i 表示该本书籍对应的序列权重，FD_i 表示该本书籍对应的时间权重，$k_i^j\in\{0,1\}$ 表示该记录是否隶属于该索引分类，$k_i^j=0$ 表示该记录不隶属于该索引分类，$k_i^j=1$ 表示该记录隶属于该索引分类。$r_j\in\{1,2,3\}$ 表示该特征词的先天重要系数，分别表示特征词在三种途径中出现的次数是 1、2、3 次。l_p 表示该记录与该索引分类相差的层数。$h\in[0,1]$ 反映每条记录对各级索引分类的影响度，通常取 1。

3.3 个性化书籍推荐模型

通过上述模型可以计算出读者特征词库中每个特征词的权重以及每个索引分类的权重，对于任一本书都包括该书籍的名称以及书籍的索引号，通过书籍名称与读者特征词库的相似度以及该书籍索引号与读者索引库的相似度的计算，得到该本书籍与读者特征库的相似度，相似度越高表示读者对该本书籍感兴趣的可能性越大，系统越应该推荐该书籍。

(1)书籍与特征词库的相似度计算

一本书籍 d 包括书籍的名称、简介、书籍的索引号等信息，考虑该书籍与读者特征词库的相似度，书籍的特征词用向量空间表示，书籍 d 和读者特征词库 W 间的相似度 sim(W,d) 可通过两向量的距离来度量，常见的有内积、余弦系数、Jaccard 系数。在此用余弦系数表示[15]：

$$sim(W,d)=cos(W,d)=\frac{W\times DW}{|W|\times|DW|}=\frac{\sum_t W_{f_k}\times DW_k}{\sqrt{\sum_t {W_{f_k}}^2}\times\sqrt{\sum_t DW_k^2}} \tag{6}$$

其中，sim(W,d) 表示书籍 d 和读者特征词库 W 间的相似度，DW 表示该书籍 d 的特征词向量，DW_k 表示该书籍 d 的某一个特征词在整个向量中的权重，W 表示读者特征词库的特征词向量，W_{f_k} 表示读者特征词库中的某一个特征词在整个向量中的权重。分子表示该书籍 d 与读者特征词库 W 相同的所有特征的权重乘积和。

(2)书籍与索引库的相似度计算

考虑一本书的索引号与读者索引库的相似度，书

籍产生的索引分类也用向量空间表示，书籍 d 和读者索引库 P 间的相似度 sim(P,d)通过两向量的距离来度量，在此使用余弦系数表示：

$$sim(P,d) = cos(P,d) = \frac{P \times DP}{|P| \times |DP|} = \frac{\sum_t P_{f_k} \times DP_k}{\sqrt{\sum_t P_{f_k}^2} \times \sqrt{\sum_t DP_k^2}} \quad (7)$$

其中，sim(P,d)表示书籍 d 和读者索引库 P 间的相似度，DP 表示该书籍 d 的特征词向量，DP_k 表示该书籍 d 的某一个特征词在整个向量中的权重，P 表示读者索引库的索引分类的向量，P_{f_k} 表示读者索引库中的某一个索引分类在整个向量中的权重。分子表示该书籍 d 与读者索引库中相同的所有特征的权重乘积的和。

(3)书籍与读者特征库的相似度计算

考虑该书籍 d 与读者特征库 I^i 之间的相似度 sim(d,I^i)，利用公式(6)和公式(7)可得：

$$sim(d,I^i) = \alpha sim(W,d) + \beta sim(P,d) \quad (8)$$

其中，sim(d,I^i)表示考虑该书籍 d 与读者特征库 I^i 之间的相似度，sim(W,d)表示书籍 d 和读者特征词库 W 间的相似度，sim(P,d)表示书籍 d 和读者索引库 P 间的相似度。α,β 为常数，根据书籍与特征词库以及索引库之间的重要关系确定。

4　实验说明

为了对上述模型的有效性进行验证，根据该模型实现了一个对应的完整系统。实验的硬件平台为：CPU Intel(R) Core(TM)2 Duo CPU T6600 @ 2.20GHz (2252 Mhz)，内存 2.00 GB (Hynix PC3 - 8500 DDR3 SDRAM SO - DIMM 1067MHz)。软件平台为：Windows 7、SQL Server 2005、JDK1.6、NetBeans IDE 7.1。实验应用上述推荐模型对图书馆相关数据进行分析，得出推荐结果。

4.1　实例读者借阅记录、学科专业、当前课程说明

(1)所有的读者借阅记录数据实例都是以表结构的形式存储在 SQL Server 2005 数据库中。为了实例化该模型，本文选取一位计算机科学与技术专业学生的借阅记录，其中实例数据所在表为 Lend，选取读者的借阅 ID 号、借书日期、还书日期、书籍索引号、借阅书籍名称等字段，该名学生的借阅记录如表 1 所示。

(2)考虑读者的学科专业。该读者是计算机科学与技术专业，《实践教学规范》和《高等学校计算机科学与技术专业公共核心知识体系与课程》选取了属于公共核心课程或者 4 个专业方向的部分示例性核心课程，在《实践教学规范》中给出了课程实验大纲。它们是：程序设计基础、数据结构、操作系统、编译原理、计算机图形学、人工智能、软件工程、数据库系统原理、数字逻辑、计算机组成基础、计算机体系结构、嵌入式系统、计算机网络[20]。课程实验大纲中的课程名称也是读者特征词库提取的潜在分析对象。

表 1　实例读者的借阅记录

ID	借书日期	还书日期	书籍索引号	借阅书籍名称
1	2009 - 9 - 26	2009 - 10 - 21	H314.3/31	大学英语常考词组、句型例解
2	2009 - 9 - 26	2009 - 10 - 21	H314.3/39	大学英语高分必背四级新短语
3	2009 - 10 - 21	2009 - 11 - 10	H319.6/332	大学英语四级考试最新真题解析 + 全真模拟解析
4	2010 - 3 - 21	2010 - 3 - 28	TP391.41/1341	新编 Photoshop CS3 中文版入门与提高
5	2010 - 3 - 21	2010 - 4 - 15	TP391.41/1548	Flash 8 中文版全程自学手册：视频教程版
6	2010 - 3 - 21	2010 - 4 - 15	TP393.092/741	征服 Dreamweaver CS4 中文版完全实战学习手册：多媒体超值版
7	2010 - 3 - 28	2010 - 5 - 7	TP312/1943	C#从入门到实践
8	2010 - 4 - 9	2010 - 5 - 11	TP312/703	Visual C#. NET 应用精彩 50 例
9	2010 - 5 - 7	2010 - 5 - 28	TP312/350	Visual C#. NET 开发实践
10	2010 - 5 - 28	2010 - 6 - 30	TP391.41/1385	中文版 Photoshop CS3 从入门到精通
11	2010 - 5 - 28	2010 - 7 - 1	TP393.092/632	Photoshop CS3 + Flash CS3 + Dreamweaver CS3 商业网站开发从入门到精通

(3)考虑读者的当前课程。当前学期读者的课程包括大学英语、C#课程设计、数据结构、操作系统、数据库系统原理等主要课程。

由此可见，从三种途径获取读者的特征词时会有彼此相同的情况，读者的特征词可能出现在借阅的书籍、读者当前学期的课程及读者专业学科三种途径的一种或几种，例如 Photoshop 出现一次，大学英语、数据结构出现两次。根据特征词在三种途径中出现的次数，确定特征词的先天重要系数 $r_j \in \{1,2,3\}$，分别表示特征词出现 1、2、3 次。

4.2　特征词库和索引库的建立和权重的计算

(1)特征词库中特征词的提取

根据读者借阅的书籍名称、学科专业核心课程、当前所学主要课程等提取特征词，读者借阅的每本书的书名以及学科专业核心课程名称、当前所学主要课程名称都可以分解为一个或者几个词组，而这每个词组就是潜在的特征词，本文特征词的提取采用了特有的正向迭代最细粒度切分算法，对所有的借阅书籍进行提取后得到读者的特征词库。

例如对表 1 第一条记录中借阅书籍名称应用正向

迭代最细粒度切分算法提取特征词得到的数据提取后得到{大学,英语,大学英语,词组,举行,实例}等特征词,根据此种方法对读者借阅的所有书籍、学科专业核心课程、当前所学主要课程等进行特征词的提取,即可提取得到该读者的特征词库中所有的特征词。

(2)特征词库中特征词权重的计算

根据渐进遗忘和滑动窗口相结合的更新算法计算每本书籍的借阅序列产生的权重。例如对表1中第8条记录进行计算,由于借阅特征序列的长度是7,第8条记录的借阅序列是5,假定k=1,应用公式(1)得到数据进行计算得到该本书籍由借阅序列对应的权重为8/6≈1.33。运用此方法即可计算出每本书籍由借阅序列对应的权重。

考虑到书籍借阅时间对权重的影响,每位读者每本书的借阅时间和平均借阅时间的比值即为借阅时间的权重,应用公式(2)和(3),例如表1中第8条记录对应的权重为32/20.5≈1.5610。

根据此种方法,应用公式(2)和(3)可确定表1中实例读者的借阅书籍索引特征 $F=(F_t,F_x,F_d)$,如表2所示:

表2 实例读者借阅书籍索引特征

书籍索引号	借阅序列	借阅时间
H314.3/31	1	25
H314.3/39	1	25
H319.6/332	2	20
TP391.41/1341	3	7
TP391.41/1548	3	25
TP393.092/741	3	25
TP312/1943	4	40
TP312/703	5	32
TP312/350	6	21
TP391.41/1385	7	33
TP393.092/632	7	34

对于读者提取词库 $W=(W_t,W_f)$ 中提取词 W_t 的权重按公式(4)得出读者的提取词库(部分),如表3所示:

表3 实例读者的提取词库(部分)

特征词	特征词权重
英语	0.6532
PS	6.7666
FLASH	4.1334
DW	4.1334
C#	11.4600

(3)索引库中索引分类的提取

对表1中第8条记录进行分析,《Visual C#. NET应用精彩50例》索引号是TP312/703,其中"TP312"是该书籍的索引号向上递推一级得到的索引号,"TP31"是该书籍的索引号向上递推二级得到的索引号,"TP3"是该书籍的索引号向上递推三级得到的索引号,"TP"是该书籍的索引号向上递推四级得到的索引号,这4个索引号都是该借阅记录产生的索引号,由此考虑每条借阅记录可以建立该读者对应的索引库。

(4)索引库中索引分类权重的计算

按公式(5)即可求出每个索引分类的权重。例如TP39的权重等于1.0490,遍历书中所有非叶子节点即可求出每个索引分类的权重,最后得到读者的索引库(部分),如表4所示:

表4 实例读者的索引库(部分)

索引分类	索引的权重
H319.6	0.3252
H31	0.1626
TP391.41	4.2602
TP391.4	2.1310
TP391	1.0656
TP393.09	4.1300
TP393.0	2.0650
TP393	1.0324
TP39	1.0490
TP31	2.8698
TP3	1.9594

4.3 书籍与读者特征库相似度的计算

假设一本书籍索引号为TP393.092/765和书名为《Dreamweaver CS4网页制作入门、进阶与提高》,根据公式(6)得出这本书与读者特征词库的相似度为6.4244,根据公式(7),这本书与读者索引分类库的相似度为5.6742,从而根据公式(8)可以得出该书籍与读者的特征库相似度为12.0986。据此方法即可求出和读者特征库相似度较大的几本书籍,这些书籍是按该算法计算出的读者最可能感兴趣的书籍,从而实现书籍的个性化推荐。

4.4 个性化推荐系统的建立

利用上述数学模型,本文实现了一个图书馆书籍个性化推荐系统。当读者输入学号后,系统会利用该数学模型,对读者的相关数据和学校馆藏书籍的相关信息进行分析,从而得到本系统为该用户提供的个性化推荐的书籍,如图1所示。

4.5 推荐方法有效性检验

为了验证本文提出的应用的有效性,对两个专业的100名读者进行更进一步调查,根据读者对图书的

图 1　读者输入学号后系统为该用户个性化推荐的书籍

评价，为每位读者推荐 10 本图书，采用信息检索领域广泛使用的查准率 Precision 和误判率 Fallout 来评价实验效果。查准率和误判率的定义公式如下：

$$\text{Precision} = \frac{\text{推荐成功的书籍}}{\text{推荐图书的总数}} \quad (9)$$

$$\text{Fallout} = \frac{\text{推荐失败时的书籍}}{\text{图书的总数}} \quad (10)$$

根据实验结果，在读者评价图书数目不同的情况下，查准率和误判率的比较如表 5 所示：

表 5　查准率和误判率的比较

用户数目	评价图书书目	查准率	误判率
25	100	47	25.8
25	200	58	20.9
25	300	64	18.2
25	400	72	16.6

可以看出，评价图书数目为 100 的读者群的推荐效果与评价图书数目为 400 的读者群的推荐效果差别较大，并且随着读者评价图书的数目不断增加，推荐效果越来越准确。

5　结　语

本文利用图书馆读者借阅记录中的书籍名称、书籍索引号、书籍借阅序列、书籍借阅时间、读者所在专业核心课程、读者当前主要课程等信息，建立了读者的包括特征词库和索引库的多兴趣特征库，运用余弦系数计算具体一本书籍与读者特征库的相似度，理论完整，系统容易实现，并对该模型的理论基础、实验过程及实验结论都做出了详细说明和有效性的检验。初步的实验表明该算法满足要求，同时具有容易实现和缩放性强的特点。但是也发现了该方法所存在的问题，特别是公式(8)中 α 和 β 值的相对大小还不能有效确定，此外，如果能考虑与读者同专业或者与读者借阅习惯相似的读者之间的相互影响则会更加全面合理，这将是未来研究的重点。

参考文献：

[1] Huang Y X, Bian L. A Bayesian Network and Analytic Hierarchy Process Based Personalized Recommendations for Tourist Attractions over the Internet[J]. *Expert Systems with Applications*, 2009, 36(1):933 – 943.

[2] RendaM E, Straccia U. A Personalized Collaborative Digital Library Environment: A Model and an Application [J]. *Information Processing and Management*, 2005, 41(1):5 – 21.

[3] 赵莉，魏治国，田广琴. 中小型高校图书馆开展个性化信息服务的措施[J]. *现代情报*, 2011, 31(2):81 – 84. (Zhao Li, Wei Zhiguo, Tian Guangqin. Small and Medium – Sized University Libraries Carrying out Measures for Personalized Information Service [J]. *Journal of Modern Information*, 2011, 31(2):81 – 84.)

[4] 顾朝晖，卢振波. 图书馆个性化服务中的用户个人信息隐私权保护[J]. *图书馆论坛*, 2011, 31(5):141 – 143. (Gu Chaohui, Lu Zhenbo. Study User's Personal Information Privacy Protect in Personalized Service of Library [J]. *Library Tribune*, 2011, 31(5):141 – 143.)

[5] 赵继海. 论数字图书馆个性化定制服务[J]. *中国图书馆学报*, 2001, 27(3): 63 – 65. (Zhao Jihai. On Personalized Customization Services of Digital Library[J]. *Journal of Library Science in China*, 2001, 27(3):63 – 65.)

[6] 马文峰. 数字图书馆个性化信息服务的探索[J]. *图书馆杂志*, 2003, 22(5):30 – 32. (Ma Wenfeng. The Exploration of Digital Library of Personalized Information Services[J]. *Library Journal*, 2003, 22(5):30 – 32.)

[7] 曹树金，罗春荣，马利霞. 论图书馆个性化服务的几个基本问题[J]. *大学图书馆学报*, 2005, 23(6):33 – 39. (Cao Shujin, Luo Chunrong, Ma Lixia. On the Personalized Library Services[J]. *Journal of Academic Libraries*, 2005, 23(6):33 – 39.)

[8] 姜雷，赵功群. 数字图书馆系统中的个性化服务模型[J]. *图书馆学刊*, 2011(9):66 – 68. (Jiang Lei, Zhao Gongqun. Personalized Service Model in the Digital Library System[J]. *Journal of Library Science*, 2011(9):66 – 68.)

[9] 张迎峰. 面向数字图书馆的个性化推荐算法研究[D]. 合肥：中国科学技术大学, 2011. (Zhang Yingfeng. Research on Algorithm of Personalized Recommendation in Digital Library[D]. Hefei: University of Science and Technology of China, 2011.)

[10] Chen R S, Tsai Y S, Yeh K C, et al. Using Data Mining to Provide Recommendation Service[J]. *WSEAS Transactions on Information Science and Applications*, 2008, 5(4):459 – 474.

[11] 刘建国，周涛，汪秉宏. 个性化推荐系统的研究进展[J]. *自然科*

学进展,2009,19(1):1-15.(Liu Jianguo, Zhou Tao, Wang Binghong. The Progress of Personalized Recommendation System [J]. *Progress in Natural Science*,2009,19(1):1-15.)

[12] 冯克鹏. 基于协同过滤的数字图书馆推荐系统研究[J]. *软件导刊*,2010,9(5):16-18.(Feng Kepeng. Digital Library Recommender System Based on Collaborative Filtering Algorithm [J]. *Software Guide*, 2010,9(5):16-18.)

[13] 赵晓岚, 张招杰. 数字化图书馆个性化推荐研究与实例[J]. *科技情报开发与经济*,2011,21(23):6-8.(Zhao Xiaolan, Zhang Zhaojie. Research on and Example of Digital library's Personalized Recommendation[J]. *Sci-Tech Information Development & Economy*,2011,21(23):6-8.)

[14] 赵麟. 基于最大频繁模式挖掘算法进行书目推荐系统的设计与实现[J]. *现代图书情报技术*,2010(5):23-28.(Zhao Lin. The Design and Implementation of the Bibliographic Recommendation System Based on Maximal Frequent Patterns Mining Algorithm[J]. *New Technology of Library and Information Service*,2010(5):23-28.)

[15] 商雪晶. 基于内容的相关书籍推荐技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2010.(Shang Xuejing. Research on Relevant Book Recommendation Technology Based on Content[D]. Harbin: Harbin Institute of Technology,2010.)

[16] 葛润霞. 基于内容聚类的协同过滤推荐系统研究[D]. 济南: 山东师范大学,2008.(Ge Ruixia. Research on the Collaborative Filtering Algorithm Based on the Content Clustering[D]. Jinan: Shandong Normal University,2008.)

[17] 李忠俊,周启海,帅青红. 一种基于内容和协同过滤同构化整合的推荐系统模型[J]. *计算机科学*,2009,36(12):142-145.(Li Zhongjun, Zhou Qihai, Shuai Qinghong. Recommender System Model Based on Isomorphic Integrated to Content-based and Collaborative Filtering[J]. *Computer Science*,2009,36(12):142-145.)

[18] 程光华. 融合内容过滤和协同过滤的智能推荐系统[D]. 南京: 东南大学,2010.(Cheng Guanghua. The Intelligent Recommender System Employing Content-based Filtering and Collaborative Filtering[D]. Nanjing: Southeast University,2010.)

[19] 黄翼彪. 实现 Lucene 接口的中文分词器的比较研究[J]. *科技信息*,2012(12):246-247.(Huang Yibiao. The Comparative Study of Chinese Word Segmentation of Lucene Interface[J]. *Science & Technology Information*, 2012(12):246-247.)

[20] 王志英,蒋宗礼,杨波,等. 计算机科学与技术专业实践教学体系与规范研究[J]. *中国大学教学*,2009(2):42-44.(Wang Zhiying, Jiang Zongli, Yang Bo, et al. Professional Practice of Computer Science and Technology Teaching System and Specifications[J]. *China University Teaching*,2009(2):42-44.)

[21] 宋丽哲,牛振东,宋瀚涛,等. 数字图书馆个性化服务用户模型研究[J]. *北京理工大学学报*,2005,25(1):58-62.(Song Lizhe, Niu Zhendong, Song Hantao, et al. Study on the User Profile of Personalized Service in Digital Library[J]. *Journal of Beijing Institute of Technology*,2005,25(1):58-62.)

[22] 马海兵,王兰成,肖辉,等. 基于《中国图书馆分类法》的用户兴趣建模方法[J]. *图书情报工作*,2007,51(8):65-68,116.(Ma Haibing, Wang Lancheng, Xiao Hui, et al. User Interest Modeling Based on Chinese Library Classification[J]. *Library and Information Service*,2007,51(8):65-68,116.)

(作者 E-mail:MJ.flye1@gmail.com)