DOI:CNKI: 11-2746/G2.20130328.1640.002

基于读者借阅二分网络的图书可推荐质量测度方法及个性化图书推荐服务*

李树青　徐　侠　许敏佳

摘　要　本文首先提出一种利用读者借阅行为特征来判断图书可推荐质量的思路,并结合读者图书借阅关系所形成的二分网络结构,设计了一种测度图书可推荐质量的迭代算法,从而为个性化图书推荐服务提供了良好的推荐客体。在上述研究的基础上,结合图书类别目录层次、标题语义信息的提取处理方法、基于加权 XML 模型的用户个性化模式表达方法及其权值扩散策略,提出了三种图书馆个性化图书推荐服务的形式,分别是特定主题的图书推荐服务、现有所借图书的修正型推荐服务和新书推荐服务。最后,文章对相关测试实验及其效果做了必要的说明。图 9。表 10。参考文献 13。

关键词　图书借阅网络　二分网络　个性化推荐

分类号　G202

The Measures of Books' Recommending Quality and Personalized Book Recommendation Service Based on Bipartite Network of Readers and Books' Lending Relationship

Li Shuqing, Xu Xia & Xu Minjia

ABSTRACT　This paper begins with the introduction of an idea about judging books' recommending quality through the characteristic analysis of the readers' lending behavior. Based on the bipartite network structure of reader and books' lending relationship, an iterative algorithm which can recognize the book's recommending quality is proposed. This algorithm can provide better recommendation objects for personalized book recommendation service. On the basis of the research mentioned-above, combined with the hierarchy of book catalog, the extract of semantic information of titles, the expression of personalized user profile based on weighted XML model and weight's spreading strategy, this paper also discusses three different methods of personalized book recommendation service in library, which include recommendation service of specific topic books, revised recommendation service of borrowed books, new books recommendation service. Finally, this paper reports some related experiments which show the improvement of effect. 9 figs. 10 tabs. 13 refs.

KEY WORDS　Network of book lending.　Bipartite network.　Personalized recommendation.

1　引言

为了给图书馆读者提供更满意的个性化推荐图书,我们必须在两方面做出有效工作:一是如何准确识别用户可能感兴趣的图书主题或者类别。由于读者一般没有完整直接的兴趣特征标识,所以常见方法是采用读者所借图书的主题信息来间接

*　本文系国家自然科学基金项目"基于通用加权 XML 模型的个性化用户兴趣本体研究"(项目编号:71103081)和江苏省高校自然科学研究面上资助项目"通用加权 XML 模型在便携式个性化用户兴趣本体中的表达方法研究"(项目编号:11KJB630001)的研究成果之一。

通讯作者:李树青,Email: leeshuqing@163.com

表达读者兴趣特征,然而读者并非对某种既定主题始终保持着浓厚的兴趣,有时也因为能力和专业水平的差距,大量所借图书可能并非自己最终所要图书,甚至还有可能借到一些质量并不高的相关图书。因此直接利用所借图书的主题信息来间接识别读者信息的方法亟需改进。二是如何有效识别不同图书的可推荐质量,从而为读者推荐令其满意的图书。传统的方法一般借助于借阅率或者借阅量等指标,即那些被更多人所借的图书通常也会被当前用户所借阅,然而我们认为具有较高专业知识背景的读者应该比一般随便借阅的读者对所借图书具有更高的认可能力。同时,由于读者兴趣的广泛性,使得某些专业性较强的图书一般在借阅率上极大地低于某些通俗类图书,这也给以单纯使用借阅率为主的推荐方法带来不利影响。

对上述第二个问题的解决有助于对第一个问题的解决,正是无法探知图书的真正质量,所以造成无法准确地从现有的读者借阅信息中识别出最有价值的所借图书信息,并据此得到读者的个性化兴趣特征,因此,本文从第二个问题的解决入手,对上述研究工作做出相应的探索。

2 文献回顾

关于对图书馆读者借阅信息的研究,传统方式大都集中于利用借阅量指标对读者和图书信息进行统计分析,较为复杂的方法往往借助于数据挖掘算法来实现,如使用多维关联规则挖掘模型来探讨读者借阅图书类型之间的关联关系或者读者学科专业、读者身份和图书类型之间的关联规则等[1]。这些研究往往可以提供陈述性较强的综合分析报告,然而很难直接嵌入到现有的借阅系统中以实现对特定用户的个性化推荐服务。要想更好地服务读者,我们需要对这些历史的借阅信息进行充分挖掘,并能根据用户的个性化兴趣特征为其提供包括图书在内的各种文献推荐服务[2]。有学者提出利用建立的读者多兴趣特征库来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,并据此实现一种可操作性和扩展性较好的混合推荐算法[3]。还有学者采用显性和隐性相结合方法,结合用户注册信息和历史查询信息,并和借阅信息一起完成对读者个性化兴趣特征信息的表达[4]。

随着复杂网络技术研究的不断深入,有学者开始从这个角度来对读者借阅信息进行分析,实践证明该方法更易于实现海量数据集合中的知识发现服务。如从借阅时间间隔分布和借阅图书大类两个角度对图书借阅网络进行分析[5];利用图书借阅网络得到图书共同借阅信息,对网络特征及其用户借阅行为进行分析,并提出院系知识依赖关系挖掘和图书借阅推荐系统这两个可能的应用领域[6]。

以上研究往往只针对简单的单顶点网络,而读者与图书之间借阅关系所构成的网络结构属二分网络(Bipartite Network),同时还可在该结构中增加节点或连接的权值以更为准确表达各种应用特征[7],因此合理挖掘不同应用环境下二分网络结构的特点并研究相应的知识发现方法显得尤为重要[8]。

国外已有学者对二分网络的特征做过研究,如结合演员电影关系、作者文献关系、《圣经》中单词和句子、网络系统中端对端信息等,对该网络特征进行统计分析[9]。还有些研究则更为细致地指出该网络中部分节点服从幂率分布而其他节点服从指数分布的特性[10]。我们把问题限定在图书借阅关系中,可以看出读者和图书各自构成一组彼此不同的节点群,但是所有的节点连接都发生在这两组节点之间,对于借阅信息而言,只有当一名读者借阅一本图书,才会在对应的读者节点和图书节点之间建立起有效连接。目前已有学者对图书借阅信息形成的二分网络进行状态统计和分析,并指出该种网络具有较高的平均集聚系数、较小的平均最短路径长度,表现出明显的小世界效应[11]。然而鲜有涉及如何利用读者图书二分网络特征来改进现有个性化文献服务的相关研究,有学者曾利用基于该网络的矩阵方法,汇总读者用户的兴趣特征权值,从而实现个性化文献推荐,不过该研究并没有很好地利用到二分网络特征[12]。

从已有研究来看,我们必须在两个方面做出努力:一是充分了解现有的读者图书借阅二元网络在

图书馆应用中的一些独有的特征，并从中发现优质的图书资源和读者资源，从而为读者提供有价值的个性化推荐结果；二是根据所发现的能够反映读者主要兴趣特征的优质图书信息，提出相应的个性化用户模式表达方法和推荐服务方法，以期获得更好的效果。关于个性化用户模式表达方法，我们在前期的研究中已经做了一些工作，初步的科研积累为此研究提供了必要的基础[13]。上述两点构成了本文的主要研究内容。

3　图书可推荐质量的权值识别方法

3.1　图书可推荐质量的定义

判断一本图书的质量并不是一件很容易的事情，主要原因在于存在读者自身兴趣特征和理解水平等因素的影响，主观性较强。因此，我们提出图书可推荐质量这个定义，并据此实现相应的测度方法和个性化推荐服务。

图书可推荐质量是指在同一类别或者同一主题下，一本图书在推送给特定读者用户群时所具有的受推荐程度。该指标并不等同于一般意义上的图书质量。相对于图书质量指标而言，具有较高可推荐质量的图书也应该具有专业性较强和撰写质量较高等共同特点，但更为重要的是，该类图书还要符合目标读者群的接受能力，因此相对于同一类别或者同一主题的其他图书，具有较高可推荐质量的图书更能适合目标读者阅读并更易于获得读者的满意。虽然具有较高可推荐质量的图书并不一定符合用户的个性化阅读特征，但是我们认为它们提供一个良好的推荐客体，结合用户的个性化兴趣特征，最终可以为读者用户提供质量较高的推荐图书。

由于图书可推荐质量与目标读者群特征的关系很密切，因此利用读者历史借阅信息，可以更为有效地识别不同图书的可推荐质量权值。对于特定主题领域，读者自身专业水平的高低决定并且依赖于所借图书可推荐质量的高低。经常借阅到较高可推荐质量图书的读者应该对所借图书的质量具有更高的推荐和认可能力，这部分读者应该对所借图书的知识领域有着较为准确的认识。由于不同图书主题覆盖面很广，同时读者也不可能熟悉所有专业领域，所以我们更强调在某既定主题下存在着此类相互加强的关系。

3.2　算法说明

以读者和图书分别作为节点，可以利用读者和图书的借阅关系形成二分网络结构。在此结构中，读者节点和图书节点并不相同，同时借阅关系产生的节点链接也只发生在两类节点之间（见图1）。

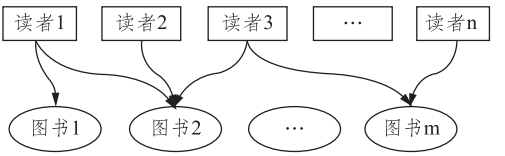

图1　读者与图书借阅关系形成的二分网络示意图

我们可以给每个读者和图书分别分配一个权值，即读者推荐能力权值 readerWeight 和图书可推荐质量权值 bookWeight。按照前文分析，图书可推荐质量和读者推荐能力互相依赖并且相互决定，图书可推荐质量权值计算方法如式1所示：

$$bookWeight_m = \sum (readerWeight_i \times repeatTime_i^m) \quad (1)$$

式1说明当前图书 i 的质量权值为借阅读者自身的质量权值 readerWeight 和重复借阅次数 repeatTime 的乘积之和。一般而言，只有对该图书主题内容有兴趣、了解该书专业相关背景知识、并且认可其质量的读者才可能重复借阅该书，所以经常被读者重复借阅的图书通常更易于受到关注。显然，读者推荐能力权值越高，重复借阅次数越多，对所借图书的可推荐质量权值影响程度就越高。

同理可以得到读者推荐能力权值的计算方法，如式2所示：

$$readerWeight_n = \sum \left(\frac{bookWeight_i}{count_i} \right) \quad (2)$$

式2中的 bookWeight 为图书可推荐质量权值，count 为当前图书的册数。在设计读者图书借阅二分网络结构时，将具有相同 MARC 号的同一种图书归并为一个节点，所以册数较多的图书通常会拥有更大的借阅量和借阅次数，而读者相对更易于借

到册数较多的图书,因此这会影响到仅通过重复借阅次数来判断图书可推荐质量的准确性。同时,利用该分散系数,还可以较好地确保迭代计算值的收敛。显然,一本图书的可推荐质量权值越大,册数越少,则对借阅读者推荐能力权值的影响程度越大。具体算法的步骤说明如下:

输入:由一组读者及其所有借阅图书组成的借阅记录集合。

输出:每个读者的推荐能力权值和每本图书的可推荐质量权值。

①给每个读者分配两个数值变量 x_r 和 y_r,其中 x_r 为读者推荐能力权值,初始值为集合中读者总数的倒数开方,y_r 用于迭代计算中保存临时数值;

②给每本图书分配两个数值变量 x_b 和 y_b,其中 x_b 为图书可推荐质量权值,初始值为集合中图书总数的倒数开方,y_b 用于迭代计算中保存临时数值;

③按照式 1 所示,设置每本图书的 y_b 值为借阅读者 x_r 值与重复借阅次数的乘积之和;

④按照式 2 所示,设置每个读者的 y_r 值为所借图书 x_r 值与册数的商之和;

⑤将每本图书的 y_b 规范化后赋予 x_b;

⑥将每个读者的 y_r 规范化后赋予 x_r;

⑦迭代计算 3 到 6 的代码。

其中,本算法中的规范化方法是使用所有权值的总和去除每个值。

4 推荐方法的实现

该方法可以从借阅历史记录中返回其中每个读者的推荐能力权值和每本图书的可推荐质量权值。据此,本文提出三种可以应用在个性化图书服务中的推荐模式,分别是特定主题的图书推荐、现有所借图书的修正型推荐和新书推荐。

4.1 特定主题的图书推荐

主要提供两种推荐方法,一种是利用图书类别,另一种是利用图书标题中的关键词。系统首先利用用户的输入信息作为图书记录的查询条件,然后在获取到的相关图书借阅集合中采用前文所述方法,从中识别具有较高可推荐质量的图书和相关读者信息。考虑到现有图书记录信息的有限性,部分图书标题信息没有很好地揭示图书主题,同时部分图书类别有时也显得过于宽泛,所以用户可以通过关键词和类别互为补充来限定推荐图书的范围。

4.2 现有所借图书的修正型推荐

读者借阅往往只根据图书名称和作者等简单信息来判断图书内容,而且现有的很多图书馆借阅系统还不能很好地提供个性化的图书推送服务。如果可以根据读者现有的所借图书信息主动推送质量更好的、完全可以替代当前所借图书的其他高质量图书,显然更能符合读者的兴趣特征。我们必须保证图书和当前读者所借图书在主题概念上非常接近,所以必须综合考虑标题语义信息和图书类别信息。具体方法如下:

(1)将文献的中文标题信息做分词处理。开始使用的是中科院中文分词包 ICTCLAS,它具有较为优异的中文分词性能和效果,然而在实际应用中发现,该方法对于图书标题之类的较为简短的中文信息处理效果并不十分理想,主要原因是对于部分词语可能存在切分不彻底的问题,甚至在不同的标题中会对同一类词语采用不同的切分方式,比如“信息检索”,如果是“信息检索原理”会切分成“信息”、“检索”和“原理”,而对于“信息检索技术”则切分成“信息”和“检索技术”。最终采用较为简单有效的二元切分方法,即将所有中文标题中每两两相邻的双字构成一个基本的二元检索单位。该方法可以保证在查询所有相关词语时具有较高的查全率,同时通过结合类别来增强查准率。

得到的所有二元检索单位在词频规律上服从齐普夫(Zipf)定律,为了测度每个二元检索单位的词语辨别力,采用的权值算法如式 3 所示:

$$twoGramWeight_i = 1 - \frac{\log(termFrequency_i)}{\log(\max(termFrequency))} \quad (3)$$

式 3 使用对数弱化部分高频词绝对数量对权值分布带来的过度两极分化的影响。

(2)获取所有和当前图书具有较高标题信息相似度并且具有一致上级类别的图书。其中标题信息相似度采用的是两两标题中共有的二元检索单位的辨别力权值和,据此可以得到按照标题相似度从高到低排序的图书记录集合。上级类别识别方法主要考虑图书分类法的特点,基本策略就是去除诸如":"、"/"、"-"和"="等符号及后缀字符,尽可能完整地保留反映图书基本类别的主要上级目录信息。实验证明此处存在误判的可能性,因此增加了调节参数,以控制最终用于计算的集合记录个数,该值越小,所获取的图书标题相似度越高,但是漏检的可能性也随之增大,因此用户可以自行调节该参数来获取对特定领域图书的推荐结果。

(3)在上述图书记录集合中,利用3.2算法从中识别出对当前图书的修正型推荐书目名单。

4.3　新书推荐

通过对读者兴趣特征的全面准确了解,可以定期或按照所需实现新书的推荐服务,推荐对象可以是新书,也可以是一些读者尚未借阅的已有图书。为此,必须准确获取和表达读者的兴趣特征。

我们采用了分类号与标题语义信息结合的读者兴趣模式表达方法,具体采用加权XML数据模型,其中每个节点使用对应的中图分类号来表示,并且上下节点关系正好对应分类法的层次关系,每个节点权值表征读者对该分类的兴趣程度。这里有两个关键问题需要解决。

(1)由于中图分类法是一个层次结构,上层节点分类号通常数量较少,但含义较为综合,多为宽泛和综合的概念,而下层分类号通常专指性较强,每本图书通常都具有较为明确的专指性分类号,因此数量太多。比较合理的选择是位于分类层次中间的类别节点,它们既能较为明确地给出具体兴趣特征,在数量上也可以很好地控制。为此,从任何一本当前读者所借图书的分类号对应节点开始,采用从下往上的权值扩散策略,扩散系数设定为0.5,最终完成对当前读者的兴趣模式表达设计。如对于读者"A023",部分兴趣模式特征见图2。

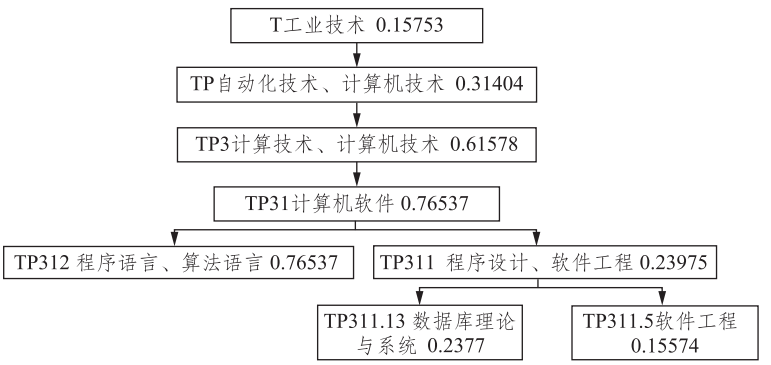

图2　基于加权XML模型、使用中图分类法层次目录结构表达的读者个性化兴趣模式部分示例

每个节点的前面内容为类目说明,后面内容为扩散后得到的兴趣度权值。经过扩散后的权值,通常会呈现以下特点:应读者主要兴趣特征的中间类目节点权值较高而两头类目节点权值较小。

(2)每个分类下的图书主题非常多,因此还需结合该类目下的各种所借图书标题信息来确定读者最感兴趣的相关主题,此时可以采用前文所述的二元分词法,也可以采用ICTCLAS分词包。按照既定的分词法,可以对每个分类节点累加其中每本图书各个关键词的辨别力权值,从而得到各个类别下最能反映读者主要兴趣的关键词。

结合上述方法,实现了新书推荐服务,该方法综合考虑新书类别和标题语义信息两个内容,基本方法如式4所示:

$$similarity_{useri}^{bookj} = \sum_{k} classWeight_{useri} \times wordWeight_{k}^{bookj} \tag{4}$$

其中,classWeight为当前用户每个类目节点对

应的兴趣度权值,wordWeight 为每本图书标题中的关键词。最终以相似度降序输出新图书列表,并依次作为当前用户的推荐新书。

5 实验说明

5.1 实验环境准备

为了对上述方法的有效性进行验证,笔者使用 JSP 实现了一个完整的图书馆个性化推荐网络平台,以便于进行数据测试和用户测试,网址为 http://www.njmars.net:8088/libs。该应用系统的硬件平台环境为:CPU Intel Xeon E5620,内存 2.4GHz 1.00GB。软件平台为:操作系统 Windows Server 2003,SQL Server 2005,JDK1.6,Eclipse 3.3。

对于实验所需要的数据,我们获取了南京财经大学图书馆汇文借阅系统 1999 年 11 月到 2010 年 9 月共计 3,968,305 条有效借阅记录,其中图书有 163,947 本,读者 70,640 名,包括学校教师和数届学生。该系统数据以 MARC 来表达同一种图书,同一种图书的每一本都具有唯一的资源号,每个读者也具有唯一的用户号。

我们采用人工评价方法来测试该方法的实际有效性和用户满意度。选择 10 位教师和 10 位学生用户,由不同测试者独立使用该系统。首先要求每位用户按照要求获得推荐图书结果,然后依次对获得的前 1 个到前 5 个具有最高可推荐质量权值的图书分别进行评价,每个结果评价层次为 3 个级别,3 为最优,1 为最差。最后统计用户对上述三个推荐功能的评价情况。

5.2 特定主题的图书推荐实验结果

我们对常见的几种图书类别和关键词进行查询测试,实验表明对于某主题领域下可推荐质量较高的图书而言,本文所设计的方法更易于识别出专业性较强的图书,尤其在某个特定的主题领域所涉及的图书总数不多的情况下。如选择“信息检索”作为关键词查询,按照借阅量排序得到图书信息(见表 1)。

表 1 借阅量排名前 10 位的“信息检索”相关图书信息

图书名称	作者	借阅量
信息检索技术	孙建军, 成颖, 丁芹编著	32
信息检索	主编张惠惠	25
看不见的网站——Internet 专业信息检索教程.1 版	(美)谢尔曼(Sherman,C.)著	23
信息检索(多媒体)教程	沈固朝主编	23
Internet 通用搜索引擎检索指南(第二版)	R.霍克	22
信息检索:从手工到联机、光盘、因特网	陆建平著	16
网络信息检索原理与技术	主编张明珍	15
信息检索	陈雅芝等编著	14
信息检索	焦玉英等编著	13
信息检索原理与方法教程	赵岩碧主编	13

按照文中所述算法得到的具有最高可推荐质量权值的图书信息(见表 2)。从表 2 中可以看出,专业性较强的图书具有非常高的可推荐质量权值,这些图书显然在读者与图书的可推荐质量权值迭代计算中不断获得更为明显的数值增长优势。

我们还可以从读者角度来对数据进行分析。选择“FoxPro”作为查询词,在按照推荐能力权值得到的读者排名中,与数据库学科关系密切的相关学科(此处包含信管和计算机两个专业)读者在前 20 位中占到 13 个,而在按照借阅量得到的读者中,该数值只有 9 个,而且排名前两位的都不属于相关专业,具体数据见表 3。相关系统截图见图 3。

表 2　可推荐质量权值排名前 10 位的“信息检索”相关图书信息

图书名称	作者	可推荐质量权值
信息检索理论与技术	主编苏新宁	9.6001774252164052E－2
现代信息检索	Ricardo Baeza-Yates 等著	6.7320889832237218E－2
信息检索(多媒体)教程	沈固朝主编	0.06565177731538456
信息检索技术	孙建军，成颖，丁芹编著	5.6856357893009704E－2
信息检索	陈雅芝等编著	5.5264467171442906E－2
文理信息检索	主编郑章飞，陈希	4.7067658296618617E－2
信息检索问题集萃与实用案例	曹志梅，范亚芳，蒲筱哥编著	4.6587994867540797E－2
信息检索原理与技术	夏立新，金燕，方志等编著	4.1625386459197208E－2
信息检索:理论与方法	叶鹰主编	3.5395401081774885E－2
信息检索	主编阎维兰，刘二稳	3.2386131155877858E－2

表 3　借阅“Foxpro”相关图书的前 20 位读者排名对比

按照推荐能力权值得到的读者排名		按照借阅量得到的读者排名	
所在部门	推荐能力权值	所在部门	借阅数
信管 041	5.6607800436050361E－3	金融 034	27
信管 041	5.0689201355924515E－3	营销与物流管理学院	25
信管 061	4.2164038966469924E－3	计算机 043	20
信息 03A	3.168926656832161E－3	物流 05A	16
广告 072	2.931158250214522E－3	工程 032	15
信息 031	2.7900434565826839E－3	计算机 014	15
信管 042	2.6648347092158094E－3	审计 043	15
信息 022	2.5366563916145615E－3	成教会计本 055	15
信管 061	2.4028357733937074E－3	计算机系	15
南京财经大学会计系	2.3173874775203533E－3	商务 032	15
自考 03 级国际商务 C 班	2.2463858506685916E－3	食品工程 001	15
信息 032	2.2421717012295187E－3	南京财经大学会计系	15
信管 051	2.2366176201245938E－3	信管 041	14
营销与物流管理学院	2.2001027220372583E－3	食品工程 051	14
食品工程 001	1.9591942001469216E－3	计算机科学与技术 001	13
食品工程 051	1.9140355060201 38E－3	信管 042	13
信管 052	1.8792790301885203E－3	信管 052	13
信管 072	1.7407284249326639E－3	信管本 011	13
社工 041	0.001717220622922	信息 03B	13
信管 062	1.6943538889475386E－3	会计 04F	13

注:为方便比较和判断,本表数据只保留专业学生数据,没有包含教师数据。

图 3 “信息检索”主题相关的图书推荐结果截图

在用户满意度实验中,每位用户被要求输入感兴趣并且熟悉的相关学科领域的 5 个关键词,同时也要求用户对现有图书馆借阅系统中前 1 个到前 5 个同一关键词的返回结果进行评价,最终统计结果如表 4 所示。

表 4 特定主题图书推荐实验的用户评价结果

对按照可推荐质量排序的图书查询结果评价(A 组)			对图书馆借阅系统常规排序的图书查询结果评价(B 组)		
前 1 个	3	67	前 1 个	3	14
	2	20		2	33
	1	13		1	53
前 2 个	3	65	前 2 个	3	13
	2	20		2	35
	1	15		1	52
前 3 个	3	60	前 3 个	3	17
	2	23		2	30
	1	17		1	53
前 4 个	3	56	前 4 个	3	20
	2	27		2	39
	1	17		1	41
前 5 个	3	55	前 5 个	3	24
	2	30		2	45
	1	15		1	31

我们对每组前 n 个数据得分值为 3 和 2 的平均值进行统计,从中可以发现利用可推荐质量进行排序输出得到的图书查询结果更令用户满意。该满意度并没有受到查询结果数量增多的影响,而图书馆借阅系统常规排序的图书查询结果则较为不理想,只是随着查询结果的增多,自然获得越来越多的满意图书,因此平均值曲线呈现逐渐上升的趋势,具体情况如图 4 所示。

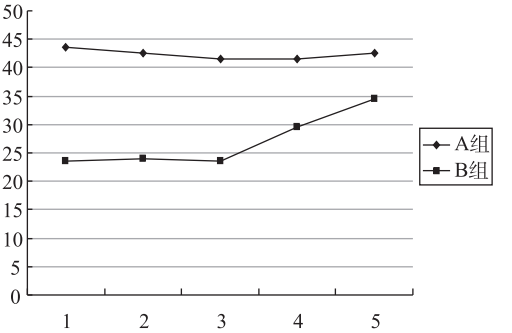

图 4 用户对每组结果的满意情况对比分析

5.3 现有所借图书的修正型推荐实验结果

如对“搜索引擎”相关图书而言,我们得到 5 条初始图书记录(见表 5)。

表 5　现有 5 本“搜索引擎”相关图书

类别号	标题	作者
F713.36/293	搜索引擎广告:网络营销的成功之路	(美) Catherine Seda 著
G354.4/28	搜索引擎优化	(美) Jennifer Grappone, Gradiva Couzin 著
G354.4/30	搜索引擎优化高级编程:PHP 版	(美)西若威齐(Sirovich,J.)著
G252.7/42	搜索引擎与信息获取技术	徐宝文,张卫丰著
G354.4/29	搜索引擎原理、实践与应用	卢亮, 张博文编著

如果选中第一本《搜索引擎广告:网络营销的成功之路》,并将调节参数设定为 8,即表示只取标题相似度最高的前 8 本图书,得到的修正型推荐结果皆与搜索引擎和广告营销密切相关(见图 5)。

而如果选中第四本《搜索引擎与信息获取技术》,则获得的修正型推荐结果和前者完全不一样,更侧重于搜索引擎技术方面的介绍,结果见图 6。

图 5　与搜索引擎和广告营销密切相关“搜索引擎”推荐结果截图

图 6　与搜索引擎技术密切相关“搜索引擎”推荐结果截图

在用户满意度实验中,我们要求每个用户选择自己所借的10本图书,评价系统返回的修正型推荐,最终统计得到所有用户对不同数量返回结果的满意度平均统计结果(见表6)。

表6 现有所借图书的修正型推荐实验的用户评价结果

组号(前i个)	评价等级(level)	得票数(vote)
前1个	3	159
	2	29
	1	12
前2个	3	174
	2	25
	1	1
前3个	3	169
	2	20
	1	11
前4个	3	161
	2	23
	1	16
前5个	3	160
	2	23
	1	17

我们对每组前n个数据的得分值按照得票数加权平均,方法如式5所示:

$$satisfactoryValue_i = \sum \frac{level \times vote}{\sum vote} \quad (5)$$

从中可以发现用户的满意度随着返回数量的增多呈现先上升后下降的趋势,总体平均值2.766大于2这个均值,说明用户满意程度较高,具体情况见图7。

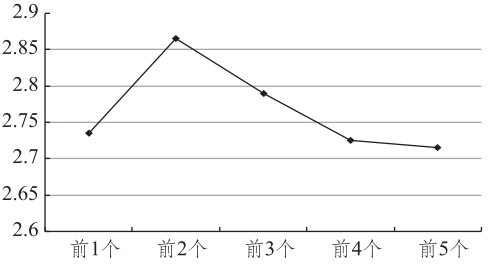

图7 用户对每组结果的满意度平均值分布

5.4 新书推荐实验结果

我们对用户进行了个性化模式的表达能力测试。如"A023"读者,该读者共有226条借阅记录,部分数据见表7。

表7 读者"A023"借阅次数排名前10的图书信息

类别号	标题	作者	借阅次数
TP3-51/29:33	Oracle9i for Windows 手册	[美]Anand Adkoli,Rama Velpuri 著	4
TP316.7/113	Windows 程序设计:第五版.上下册	(美)[C.佩措尔德]Charles Petzold 著	3
TP3-51/29:31	Oracle9i PL/SQL 程序设计	(美)Scott Urman 著	3
F014.6/12	现代产业经济学	杨士朴,夏大慰主编	3
TP31/174	软件报:2001年合订本.上下册	《软件报》编辑部编	3
TP312/1150:2	Java 2 核心技术.卷Ⅱ.volume Ⅱ,高级特性,adanced features	(美) Cay S. Horstmann, Gary Cornell 著	3
H31-43/117	博士研究生英语读写教程	陈大明,徐汝舟主编	3
TP393.092/450	JSP 程序设计	Vivek Chopra, Jon Eaves 等著	3
TP311.138/378	Oracle Discoverer10g 手册:创建、维护和管理高效的即席查询	(美)史密斯(Smith,M.A.),(美)史密斯(Smith,D.A.)著	3
TP312/1360	JSP 高级程序设计	Vivek Chopra, Jon Eaves 等著	3

在按照新书推荐算法得到的相关读者兴趣模式中，主要的读者兴趣特征见表8。

其中在“TP311”类和“TP312”类下，最能反映用户兴趣特征的关键词见表9。

实验系统实现了对用户个性化模式的表达和展示功能（见图8）。

表8　读者“A023”兴趣模式中具有最高权值的前10个类别特征及其权值

类别号	类别名称	权值
TP312	程序语言、算法语言	1
TP31	计算机软件	0.765368852
TP3	计算技术、计算机技术	0.615778689
TP393.092	网络浏览器	0.557377049
TP393.09	计算机网络应用程序	0.475409836
TP311.138	数据库系统：按系统名称分	0.393442623
TP	自动化技术、计算机技术	0.314036885
TP311	程序设计、软件工程	0.239754098
TP311.13	数据库理论与系统	0.237704918
TP393.08	计算机网络安全	0.229508197

表9　与读者“A023”相关的、类别“TP311”和“TP312”下具有最高权值的前5个关键词

TP311 类		TP312 类	
关键词	权值	关键词	权值
ORACLE	6.8575951602940926	JAVA	5.9497040952520743
SQL	3.7724521394734429	UML	3.8891092182475839
SERVER	3.2173403738318864	XML	3.4792902710497424
10G	2.8818117598366442	VISUAL	3.1347764579322615
ECLIPSE	2.8041789739508225	C	2.9177404593398966

图8　读者“A023”用户个性化模式的效果展示

在用户满意度实验中,我们在200名学生用户中进行满意度评价测试,首先将2011年以来的图书做成新书列表,并计算每本图书与读者个性化模式之间的相似度,并降序输出前5本相似度最高的新书作为推荐图书。每位用户被要求对每本推荐图书进行满意度评价,结果见表10。

表10 新书推荐实验的用户评价结果

组号(前i个)	评价等级(level)	得票数(vote)
前1个	3	189
	2	10
	1	1
前2个	3	170
	2	15
	1	15
前3个	3	184
	2	14
	1	2
前4个	3	190
	2	6
	1	4
前5个	3	179
	2	10
	1	11

我们对每组前n个数据的得分值按照式5计算用户满意度,结果见图9。

其中,读者对新书推荐列表中的图书,并没有呈现较为明显的随着推荐数量增加而满意度逐渐增高或者降低等现象,同时全部满意度的平均值为2.879,显示出用户非常满意最终的推荐结果。

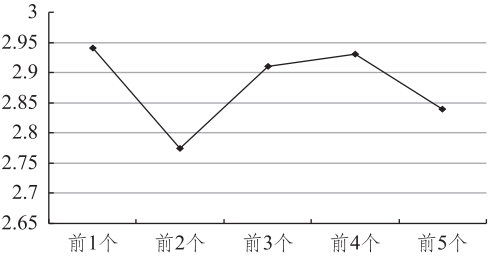

图9 用户对每组结果的满意度平均值分布

6 总结与展望

本文提出了基于读者图书借阅二分网络结构的测度图书可推荐质量的方法,并据此完成相关的个性化图书推荐实验,初步实现预期的设计目标。不过该方法存在很多需要进一步研究和改进的地方,主要有以下两点:一是现有的语义分析和语义识别能力尚需完善,这样可以更好地提高查准率;二是我们注意到时间维度在表达读者兴趣特征时的重要性和可利用性,如何有效地结合时间信息来进一步提高对用户个性化模式表达的准确度,是下一阶段主要的研究目标。

参考文献

[1] 王家胜,牟肖光. 读者借阅多维关联规则挖掘模型的建立与分析[J]. 计算机应用, 2011, 31(11): 3084-3086. (Wang Jiasheng, Mou Xiaoguang. Establishment and analysis of the mining model of multidimensional association rules of reader loan[J]. Journal of Computer Applications, 2011, 31(11): 3084-3086.)

[2] 胡蓓蓓. 基于知识决策的数字图书馆个性化推荐[J]. 情报学报, 2007, 26(3): 448-455. (Hu Beibei. Personalized recommendation in digital library based on knowledge decision-making[J]. Journal of the China Society for Scientific and Technical Information, 2007, 26(3): 448-455.)

[3] 马健,杜泽宇,李树青. 基于多兴趣特征分析的图书馆个性化图书推荐方法[J]. 现代图书情报技术, 2012, 28(6): 1-8. (Ma Jian, Du Zeyu, Li Shuqing. Personalized book recommendation algorithm based on multi-interest analysis in library[J]. New Technology of Library and Information Service, 2012, 28(6): 1-8.)

[4] 张炜,李斌. 基于联机公共查询目录的读者行为挖掘的个性化智能服务系统构建[J]. 情报理论与实践, 2009, 32

(10): 68 - 71. (Zhang Wei, Li Bin. Construction of the individual intelligent service system based on the mining of readers' behavior in OPAC database[J]. Information Studies: Theory & Application, 2009, 32(10): 68 - 71.)

[5] 王福生,杨洪勇. 图书管理系统中的借阅行为分析[J]. 复杂系统与复杂性科学, 2012, 9(1): 55 - 58. (Wang Fusheng, Yang Hongyong. Books-borrowing behavior in library management system[J]. Complex Systems and Complexity Science, 2012, 9(1): 55 - 58.)

[6] 燕飞,张铭,孙韬,等. 基于网络特征的用户图书借阅行为分析——以北京大学图书馆为例[J]. 情报学报, 2011, 30(8): 875 - 882. (Yan Fei, Zhang Ming, Sun Tao, et al. An analysis of network-based users' book-loan behavior: A case study of Peking University library[J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(8): 875 - 882.)

[7] Morris S A, Yen G G. Construction of bipartite and unipartite weighted networks from collections of journal papers[EB/OL]. [2012 - 11 - 01]. http://arxiv.org/abs/physics/0503061.

[8] 傅林华,郭建峰,朱建阳. 图书馆图书借阅系统与单标度二元网络模型[J]. 情报学报, 2004, 23(5): 571 - 575. (Fu Linhua, Guo Jianfeng, Zhu Jianyang. Borrowing and reading system of books in library and the single-scale bipartite networks model[J]. Journal of the China Society for Scientific and Technical Information, 2004, 23(5): 571 - 575.)

[9] Latapy M, Magnien C, Vecchio N D. Basic notions for the analysis of large two-mode networks [J]. Social Networks, 2008, 30(1): 31 - 48.

[10] He Y H, Tian L X, Huang Y. A simple dissortative bipartite network model[J]. International Journal of Nonlinear Science, 2012, 13(2): 200 - 203.

[11] 王进良,张鹏,狄增如,等. 北京师范大学图书借阅系统的网络分析[J]. 情报学报, 2009, 28(1): 137 - 141. (Wang Jinliang, Zhang Peng, Di Zengru, et al. Network analysis based on loan system of library of Beijing Normal University [J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(1): 137 - 141.)

[12] Li Nannan, Zhang Ning. The bipartite network study of the library book lending system[J]. Complex Sciences, 2009, 4(1): 773 - 782.

[13] 李树青. 基于加权 XML 数据模型的个性化本体研究[J]. 情报学报, 2010, 29(10): 826 - 834. (Li Shuqing. Study of personalized ontology based on weighted XML data model[J]. Journal of the China Society for Scientific and Technical Information, 2010, 29(10): 826 - 834.)

李树青　南京财经大学信息工程学院信息管理与信息系统系主任,副教授,硕士生导师。
通讯地址:江苏省南京市仙林大学城南京财经大学信息工程学院。邮编:210046。

徐　侠　南京人口管理干部学院工商管理系副教授。
通讯地址:江苏省南京市龙蟠路 177 号南京人口管理干部学院工商管理系。邮编:210042。

许敏佳　南京财经大学图书馆技术部主任,工程师。
通讯地址:江苏省南京市仙林大学城南京财经大学图书馆技术部。邮编:210046。

(收稿日期:2012 - 11 - 26;修回日期:2013 - 01 - 23)