

数字图书资源聚合及精准化推荐方法研究^{*}

吴彦文¹ 牛晓璇¹ 胡炎贵¹ 王馨悦² 何秀玲³

(1. 华中师范大学物理科学与技术学院, 武汉 430079; 2. 华中师范大学信息管理学院, 武汉 430079;

3. 华中师范大学国家数字化学习工程技术研究中心, 武汉 430079)

摘要: 针对数字资源信息过载、信息异构、资源推荐效果不理想等问题, 改进传统数字资源聚合模型和相似度计算方式, 本文提出一种数字资源聚合模型, 并融合协同过滤的推荐思想, 利用该聚合模型进行相似度计算得出资源和用户的近邻集合, 基于此设计精准化资源推荐算法, 最后以馆藏图书资源为例对模型进行验证。结果表明, 本文构建的方法能够对数字资源进行有效聚合, 并挖掘图书的语义信息, 同时结合用户兴趣模型, 为用户提供精准化的资源推荐。

关键词: 数字资源聚合; 作者耦合; 语义网; 协同过滤

中图分类号: G250

DOI: 10.3772/j.issn.1673-2286.2018.11.002

大数据时代, 海量数字资源的爆炸增长已经成为不可逆转的趋势, 数字资源在丰富传统纸媒资源的同时, 也带来诸多问题。如由于表现形式不同, 不同类型资源可能包含相同信息, 用户在检索时无法获取所需信息; 由于信息处理能力有限, 用户无法对包含海量信息的资源进行有效整合、组织及内化; 面对海量数字资源, 用户在检索所需信息时势必会浪费大量精力与时间, 产生“信息迷航”。诸多问题成为网络信息组织与检索的障碍, 也在一定程度上降低了数字资源的附加价值及用户吸引力。近年来, 国内外学者将这一问题的解决途径聚焦在数字资源的重组及精准化推荐上。然而, 目前的数字推荐系统所使用的推荐方法角度较单一, 且缺少多维度的图书资源聚合与推荐。因此, 如何有效地基于数字图书资源进行精准化推荐, 满足各类用户多元化的知识需求, 成为当下研究的热点。

相关研究显示, 资源聚合能够发现资源间潜在的联系, 通过系统整合、有效聚合及深度挖掘数字资源语义信息的途径, 形成一体化的知识聚合网络^[1]; 而精准的推荐技术能够智能化过滤冗余信息, 深入挖掘用户兴趣偏好和资源访问行为, 主动向用户进行数字资源推荐, 满足用户多元化的服务需求^[2-3]。如Selamat等^[4]通过

语义网提取数字资源概念间的关联形成树形结构, 为用户提供知识检索服务; 黄文碧^[5]基于元数据关联的数字资源聚合模型并为用户进行相关资源的推送服务; 严春子^[6]提出一种公共文化数字资源聚合服务方法, 将分散的数字资源予以整合, 从而加速用户获取资源的过程; 胡媛等^[7]从知识导航服务、语义个性化检索和信息集成推送3个层面研究数字图书馆社区集成推送服务, 达到一定的聚合与推荐效果; 毕强等^[8]通过构建本体、语义相似度及谱聚类等方法, 研究数字文献资源聚合及服务推荐的流程, 提升聚合质量与推荐效果。因此, 基于数字资源聚合进行精准化推荐已成为国内外学者关注的热点问题。但前述相关研究中, 缺乏对馆藏图书作者社会化群体信息的考虑, 且资源推荐方法(如基于内容的推荐、协同过滤推荐、基于矩阵分解等单一推荐技术)会使结果片面, 即在推荐质量上, 仍需进一步改进。

因此, 本文提出一种基于数字资源聚合模型的多标签协同过滤推荐方法: 首先, 该方法将作者耦合分析、语义网技术引入数字资源聚合模型, 以此模型作为计算依据来代替传统的相似度计算方式, 产生基于作者群体和基于相似内容的推荐资源; 其次, 结合用户兴趣模型产生基于相似用户的推荐资源; 最后, 融合3种推荐结

^{*}本研究得到教育部人文社会科学研究规划基金项目“智慧教室环境下课堂交互有效性量化研究”(编号: 17YJA880030)资助。

果,实现图书推荐。实验结果表明,该方法融合作者-资源-用户3个角度,提高了图书推荐的新颖性和精准度。

1 资源聚合与推荐相关技术

1.1 作者耦合分析

馆藏资源包含图书、报纸、期刊等多种数字资源,而这些资源共同的桥梁为作者本身。作者耦合关系通常包括作者引文耦合和作者关键词耦合等,单一选择作者引文或作者关键词无法显示某一领域知识关联网络的全貌;如果将多个耦合结合使用,则可以更好地研究一个领域的信息结构全貌^[9],并在此基础上更精确地度量作者耦合度,为用户带来相似作者群体的资源推荐,提高推荐结果的新颖度。考虑到馆藏图书中并非所有作者均有索引文献,作者引文耦合存在关系网络稀疏的局限性,因此,本文选择作者分类号耦合与作者关键词耦合构建作者耦合分析网络。

1.2 语义网技术

由于存在半结构化或非结构化数据,馆藏数字资源间缺乏相互的关联关系,使得数字资源形成一个个“孤岛”。而语义网技术可以利用语义标签将离散的数字资源连接成紧密的、结构化的知识关联网络,优化数字资源组织结构并提高个性化推荐的水平和效率。其中,语义相似度是构建语义网的关键。一方面,基于本体的概念语义相似度一般通过本体概念的内容、属性或距离进行相似度计算,但本体所包含的语义信息相对较复杂,不能充分表征其概念的语义内涵,因此计算精度不高;另一方面,基于语义词典的语义相似度计算方法一般基于完备的英文或中文语义词典,如WordNet、FrameNet、MindNet等英文词典和《知网》《同义词词林》等中文词典^[10]。考虑到基于《同义词词林》的概念语义相似度计算方法更符合人们对词汇的理解方式^[11],因此本文采用基于《同义词词林》的语义相似度计算方法构建语义网。

1.3 个性化推荐算法

数字资源聚合是对异构数字资源的重构和再组织,其中一个重要的目的是进行精准化的资源推荐,代

替用户发现其潜在需求的数字资源,根据用户兴趣偏好进行智能推荐。个性化推荐算法主要有基于内容的推荐、协同过滤推荐及组合推荐算法^[12]。

目前已经有许多学者将个性化推荐应用到数字资源推荐中。国外学者Shelton等^[13]基于内容的文本相似性对用户进行数字图书馆资源推荐,创建有效的资源检索系统;Tsuiji等^[14]基于图书馆的借阅信息和图书内容进行图书相似度计算,并采取SVM进行图书资源的推荐;Fedelucio等^[15]提出一种基于内容的跨语言推荐系统,使用维基百科和BabelNet构建基于概念的内容表示。在国内,周之诚^[16]基于用户意图聚类进行数字资源推荐,其本质上使用的是基于用户的协同过滤推荐思想;周玲元等^[17]融合用户情境改进了用户相似度计算,提出基于情景感知的协同过滤推荐算法;曾子明等^[18]利用用户动态的兴趣变化对用户-资源评分矩阵进行数据填充,并结合基于用户的协同过滤算法进行个性化推荐。

由此可见,①在数字资源推荐系统中推荐技术大多为基于内容、基于用户或项目的协同过滤的方法,推荐的资源角度较单一;②仅按照图书文本相似性进行资源推荐会使推荐结果越来越狭窄,难以给予用户新颖性的阅读感受;③推荐技术对于数字资源的应用,主要集中在以图书为单位进行的推荐,用户所需的知识服务并不全部存在于一本图书中,因此缺少对图书资源语义上的深层融合。

本文针对以上问题,从多个角度出发,结合基于多标签的协同过滤推荐算法,将相似作者、相似内容及相似用户多角度的数字资源同时考虑,并借助资源聚合模型改进传统的相似度计算方式,为用户进行精准化的资源推荐。

2 数字资源聚合及推荐算法流程

为提高资源聚合度和推荐准确性,本文设计了数字资源聚合及精准化推荐框架(见图1),具体步骤如下。

(1) 基于作者耦合分析计算作者间的关联度,构建作者关联度矩阵;

(2) 基于语义网计算资源间的语义相似度,构建资源语义相似度矩阵;

(3) 通过用户历史行为构建用户兴趣偏好模型,缓解数据集的稀疏性并计算用户相似度以构建用户相似度矩阵;

(4) 结合协同过滤和社会化网络思想,利用作者

关联度矩阵、资源相似度矩阵及用户相似度矩阵产生相似作者资源、相似内容资源及相似用户资源3个标签的推荐资源候选集合;

(5) 按照加权计算的方法融合3个标签的推荐结果, 降序排列后选取前 N 个资源推荐给目标用户。

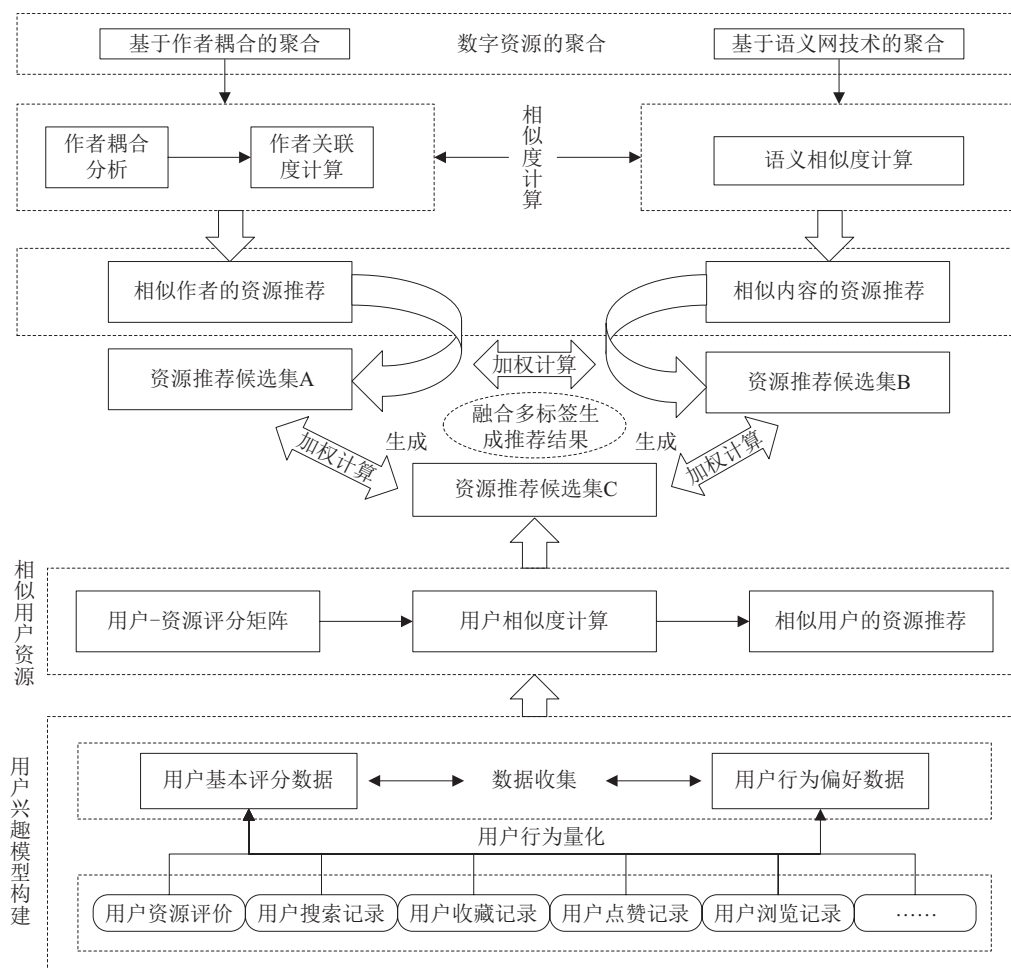


图1 数字资源聚合及精准化推荐框架

2.1 作者关联度计算

借鉴学术论文的作者分类号耦合和作者关键词耦合, 本文提出基于数字图书资源的作者耦合分析方法: 使用2位作者分类号集合中相同的数量来决定作者分类号的耦合强度, 而作者关键词集合中相同关键词的数量决定了作者关键词耦合强度^[19]。由于每位作者的图书远不及论文发表的数量, 单纯按照作者分类号的计量方法会造成网络稀疏, 因此本文改进文献[19]中的作者分类号方法, 引入分类号级别权重。

(1) 作者分类号耦合强度计算。给定作者属性集合, 分类号集合为 $C=\{c_1, c_2, c_3, \dots, c_q\}$, 关键词集合为 $K=\{k_1, k_2, k_3, \dots, k_r\}$, 采用VBA自建程序建立作者 i 和

作者 j 的耦合矩阵 $C_{m \times m}$ 。基于此, 本文提出基于作者耦合强度的作者关联度计算方法见公式(1)。

$$C_{ij} = \sum_{n=1}^m \lambda_n f(c_{ij}) \quad (c_{ij} \in \{C_i\} \cap \{C_j\}) \quad (1)$$

其中, C_{ij} 代表作者 i 和作者 j 的分类号耦合强度; n 代表作者图书分类号最高相同级数, 取值范围为 $\{1, 2, 3, 4, 5\}$, 分别对应一级至五级分类号; λ_n 取值范围为 $\{0.2, 0.6, 0.8, 0.96, 1\}$, 其值分别对应分类号最高相同级数的权重; $f(c_{ij})$ 是作者图书相应级数中相同的分类号数量。例如, 作者 i 的分类号集合为 $\{G02, G03, G07, TP23\}$, 作者 j 的分类号集合为 $\{G02, G03, G235, TP213\}$, 传统分类号耦合强度计算得出为2; 对不足五级的分类号用“0”补足以便于计算, 计算得到改进后的

耦合强度为 $1 \times 2 + 0.2 \times 1 + 0.8 \times 1 = 3$ ，直观上可以看出耦合强度量化更符合实际情况。

(2) 消除分类号的规模差异影响。为消除作者的分类型、关键词规模差异所带来的影响，采取对耦合强度进行标准化的办法解决此问题。Salton标准化公式见公式(2)。

$$C'_{ij} = \frac{C_{ij}}{\sqrt{C_i \times C_j}} \quad (2)$$

其中， C_i 是作者 i 集合中包含的分类号总数量， C_j 是作者 j 集合中包含的分类号总数量。

(3) 按照同样方法标准化作者关键词耦合强度，其构建方法在此不再赘述。按照加权平均的方法计算作者的关联度 S_{ij} 见公式(3)。

$$S_{ij} = pC'_{ij} + qK'_{ij}, \quad (p + q = 1) \quad (3)$$

其中， C'_{ij} 为作者分类号耦合强度， K'_{ij} 为作者关键词耦合强度， p 和 q 为权重值。

2.2 语义相似度计算

由于图书具有超高维的特点，对整本书进行关键词选取势必会大幅增加算法的时间复杂度。图书的标题、大纲等知识描述信息浓缩了整本图书的语义信息，因此本文利用图书的知识描述信息进行关键词的选取，具体构建步骤如下。

(1) 基于TF-IDF技术对图书的知识描述信息进行关键词抽取。选取词频最高且区分度最高的前 k 个关键词，并表示为《同义词词林》6层编码形式，如关键词Ba03C01和Ba03A02。

(2) 语义相似度计算。本文采用的是《同义词词林(扩展版)》，采用六层编码原则，具体见表1。

表1 词语编码表

编码层	符号表示	符号性质	编码层	符号表示	符号性质
1	大写字母	大类	4	大写字母	词群
2	小写字母	中类	5	两位数字	原子词群
3	两位数字	小类	6	=, #, @	

计算关键词的语义相似度构成语义描述。关键词集合A1和A2的基于《同义词词林(扩展版)》的语义相似度 $S(A1, A2)$ 计算见公式(4)。

$$S(A1, A2) = \frac{1}{n} \sum_{i=1}^k dis_i(A1, A2) \times \cos\left(a \times \frac{\pi}{180}\right) \times \left(\frac{a-b+1}{a}\right) \quad (4)$$

其中， $dis_i(A1, A2)$ 是由两个词的最近共同祖先所在层数决定的，参考文献[11]提出的相似度计算方法，取 $dis_i(A1, A2) \in \{0.1, 0.65, 0.8, 0.9, 0.96\}$ 其值分别对应最近共同祖先所在层数从小到大的取值。如关键词Ba03C01和Ba03A02，其最近共同祖先所在层数为3，则 $dis_i(A1, A2) = 0.8$ 。 a 是同义词词林中分支层的总节点个数； b 是两关键词所在分支间的距离。

(3) 文档匹配。将图书看作一个文档，对文档进行语义描述匹配。在判断匹配是否成功时，需要对关键词匹配和知识节点匹配取一个阈值 θ ，若大于该阈值则匹配成功。根据语义相似度匹配与之相似的图书列表构成关联数据，由此达到数字资源语义互联的目标。

2.3 基于多标签的推荐算法

2.3.1 用户兴趣模型构建

对用户进行资源推荐需要构建用户兴趣模型，本文通过用户-资源评分矩阵来表示用户的兴趣模型，具体描述见公式(5)。

$$S_{m \times n} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1n} \\ s_{21} & s_{22} & \dots & s_{2n} \\ \dots & \dots & \dots & \dots \\ s_{m1} & s_{m2} & \dots & s_{mn} \end{bmatrix}, \quad s_{ur} \in \{1, 2, 3, \dots\} \quad (5)$$

其中， $s_{ur}(u \in \{1, \dots, m\}, r \in \{1, \dots, n\})$ 是用户 u 对数字资源 r 的评分值，评分越高，用户对数字资源的偏好程度越大，反之越小。

目标用户对资源的评分较少，由此带来数据稀疏的问题，因此引入用户行为进行数据填充。本文将用户对各种图书的搜索、浏览、收藏、下载等行为量化为评分值丰富 $S_{m \times n}$ 中的数据，以5分制评分为例，其行为量化评分表见表2。

表2 用户行为量化评分表

用户行为	量化值	用户行为	量化值	用户行为	量化值
搜索	+3	收藏	+3.5	收藏+下载	+4.5
浏览	+2.5	下载	+4		

2.3.2 基于相似作者的推荐

基于作者耦合对用户进行资源推荐的基本思想：首先获取目标用户 u 已评分的图书作者，构成作者群体集合，根据2.1节所述的作者关联度降序排列，取前 k 名作

者的图书资源构成聚类中心 AA_i ，此时 AA_i 表示与目标作者群体最相似的作者所属图书资源集合；其次，对图书资源集合中的资源评分进行预测并降序排列；最后，选取前 N 个资源构成基于相似作者的推荐候选集合 A 。

2.3.3 基于相似用户的推荐

相似用户的资源推荐是基于用户的协同过滤推荐，其基本思想是计算用户相似度并构建用户相似度矩阵，选取前 k 个相似用户评分过的资源构成推荐集合，预测用户 $u \in B$ 对资源的评分，对评分进行降序排列并选取前 N 个资源推荐给用户 u 。

2.3.4 基于相似内容的推荐

相似内容的资源推荐是基于内容的协同过滤推荐，其基本思想是构建并维护资源相似度矩阵，选取前 k 个目标资源的最近邻居构成推荐集合，预测用户 u 对资源 r 的评分，对评分进行降序排列并选取前 N 个资源推荐给用户 u 。

2.4 推荐结果生成

经过上述计算，产生目标用户 u 的基于相似作者、相似用户及相似内容的3种资源推荐候选集，这3种结果分别对应馆藏资源系统中的作者、用户、资源3种标签的资源，即社会化群体数字资源、用户偏好数字资源及相似内容数字资源，形成作者-用户-资源多元化综合知识网络。这3种角度的重要程度相同，因此按照加权平均的方法对推荐候选集的预测评分进行归一化处理并加权

计算。融合3种推荐候选集合后按照评分降序排列，选取前 n 个资源作为最终的推荐结果推荐给目标用户 u 。本文最终向用户推荐前10个图书资源。

3 实验过程及结果分析

3.1 实验数据集

为验证推荐方法的有效性，选取图书馆系统后台图书评分及借阅记录数据集进行实验分析，时间跨度为2016年1月—2017年12月，经过人工浏览的方式随机选取252个用户对500本图书的评分信息，共得到711条日志信息记录，每本图书评分数值由1~5分不等。将这些数据按照8:2的比例划分为训练集和测试集，80%为训练集，20%为测试集，将数据以SQL文件的格式导入数据库。按照表2所示的用户行为量化评分表对数据集进行数据填充，计算得出填充前的数据稀疏度为93.64%，填充后的数据稀疏度为11.86%。可以看出，用户行为量化方法对数据稀疏起到一定的缓解作用。

3.2 作者耦合及关联度计算

图书的主编信息一般代表该书在作者层面上的语义信息，同时以主编作为图书作者也降低了计量和分析的难度，因此本文取主编为图书作者，提取作者分类号集合和作者关键词集合，其中作者关键词包含作者的社会身份信息、研究方向、流派等，部分结果见表3。使用VBA自建程序构建作者耦合矩阵，同时使用Python对公式(2)~(3)编程计算作者之间的关联度。部分结果见表4。

表3 数字图书资源作者信息数据集

图书编号	图书名称	作者名称	关键词属性值
1	《惯性导航》	秦永元	卡尔曼滤波、捷联惯导、空中对准、故障检测等
2	《虚拟现实技术》	黄海	计算机应用、工业通用技术及设备、博物馆文化等
3	《深入理解大数据 大数据处理与编程实现》	黄宜华	并行化、Spark、Mapreduce、分布并行计算等
4	《数字图像处理与分析》	龚声蓉	视频分割、图象分割、局部运动估算、运动对象等
5	《多尺度图像融合理论与方法》	才溪	低通滤波器、Contourlet变换、图像融合等
6	《MATLAB图像处理 程序实现与模块化仿真》	赵小川	仿生六足机器人、DSP、传感器、制度化等
7	《算法设计与分析》	王晓东	Ontology、语义web、角色概念、元模型等
8	《云计算与云应用》	万川梅	安全模型、系统集成、存储模型、可信计算等
...	...	...	...
500	《计算机网络》	谢希仁	ATM、宽带ATM、自适应滤波、无限蜂窝网等

表4 作者关联度矩阵

作者编号	1	2	3	4	...	500
1	1.000 000	0.042 854	0.072 968	0.052 378	...	0.064 200
2	0.042 854	1.000 000	0.075 262	0.047 328	...	0.125 040
3	0.072 968	0.075 262	1.000 000	0.123 775	...	0.121 490
4	0.052 378	0.047 328	0.123 775	1.000 000	...	0.249 520
5	0.063 521	0.092 470	0.046 618	0.430 824	...	0.050 830
6	0.062 560	0.080 940	0.039 639	0.434 065	...	0.119 139
7	0.055 190	0.034 415	0.076 380	0.044 620	...	0.210 310
8	0.145 512	0.030 810	0.218 729	0.040 317	...	0.223 210
...	...	...	...	...	...	...
500	0.064 200	0.125 040	0.121 490	0.249 520	...	1.000 000

3.3 语义网构建及相似度计算

每个数字资源都可以看作一个文档，文档的关键词选取不规范或不科学则会导致其难以反映图书的语义特征。本文利用TF-IDF技术对文档进行词频统计与分析，从中选取词频较高且能表征语义内涵的词汇作为

文档的关键词，并将其记录入库。利用Python语言进行编程计算资源间的语义相似度。部分计算结果见表5。

3.4 用户兴趣模型构建及相似度计算

用户相似度矩阵部分计算结果见表6。

表5 资源间语义相似度矩阵

图书编号	1	2	3	4	...	500
1	1.000 000	0.412 954	0.321 967	0.458 149	...	0.263 420
2	0.412 954	1.000 000	0.275 062	0.473 280	...	0.142 504
3	0.321 967	0.275 062	1.000 000	0.323 717	...	0.149 000
4	0.458 149	0.473 280	0.323 717	1.000 000	...	0.264 951
5	0.290 498	0.292 487	0.206 618	0.330 149	...	0.050 913
6	0.116 676	0.090 942	0.100 639	0.089 265	...	0.189 439
7	0.575 139	0.384 465	0.363 989	0.463 462	...	0.310 631
8	0.459 512	0.309 284	0.198 079	0.460 317	...	0.262 100
...	...	...	...	...	...	...
500	0.263 420	0.142 504	0.149 000	0.264 951	...	1.000 000

表6 用户相似度矩阵

用户编号	1	2	3	4	...	252
1	1.000 000	0.110 646	0.418 642	0.263 820	...	0.350 485
2	0.110 646	1.000 000	0.109 571	0.172 175	...	0.172 676
3	0.418 642	0.109 571	1.000 000	0.265 345	...	0.527 484
4	0.263 820	0.172 175	0.265 345	1.000 000	...	0.404 711
5	0.372 062	0.650 939	0.251 336	0.139 833	...	0.193 880
6	0.421 963	0.209 361	0.620 784	0.602 087	...	0.511 630
7	0.426 440	0.119 650	0.497 158	0.871 259	...	0.471 150
8	0.271 645	0.102 858	0.634 318	0.109 773	...	0.228 554
...	...	...	...	...	...	...
252	0.350 485	0.172 676	0.527 484	0.404 711	...	1.000 000

在上述得出作者关联度矩阵、资源间语义相似度矩阵及用户相似度矩阵的基础上,产生基于相似作者、相似内容及相似用户的资源推荐,融合3种标签的推荐候选集合形成最终的资源推荐结果。

3.5 推荐结果分析

为合理评估本文提出的推荐方法的准确性以及确定合理的最近邻居数目,本文使用平均绝对偏差MAE^[22]来衡量推荐结果的准确性,分别计算不同推荐算法(基于用户的协同过滤推荐、基于项目的协同过滤推荐等)的MAE值并进行比较。为方便描述,本文提出的多标签融合推荐算法记为multi-CF,基于相似用户的推荐记为U-CF,基于相似内容的推荐记为I-CF,基于相似作者的推荐记为A-CF。MAE值的比较结果见图2。

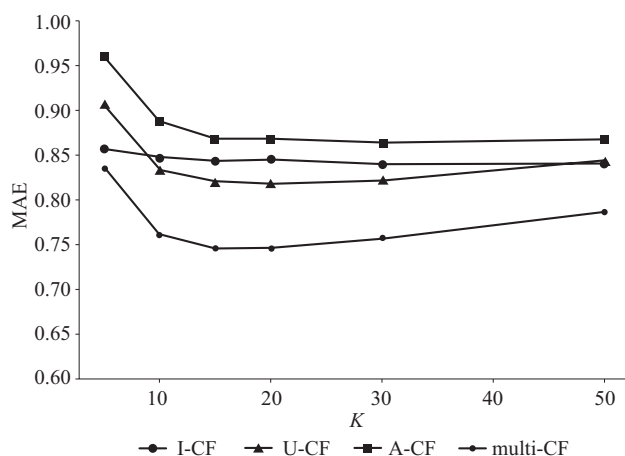


图2 不同邻居数目下4种方法的MAE值结果比较

随着最近邻居数目的增多,各个推荐算法的MAE值逐渐下降,说明选择较多的最近邻居数目可以对资源进行聚类并提高推荐算法的准确度;当最近邻居数目进一步增多时,4种推荐算法的MAE值均出现不同程度的上升,这是由于过多的最近邻居在实际聚类情况下产生了失真,某些不必要的离群点也被聚类到计算过程。可以看出,同其他3种推荐算法相比较,当最近邻居数目 $K=15$ 时,本文提出的多标签资源推荐算法的MAE值最小,这说明该方法可以有效聚合数字资源并提高算法推荐准确性,但其对 K 值较敏感,因此在推荐系统的实际应用中需要进一步选取适当的聚类数目。

另外,本文选取了查全率(Precision)、查准率(Recall)及 F 值作为评价指标,进一步探究推荐算法的准确度。本文提出的多标签推荐算法与其他3种方法

比较的结果见图3。相比较其他3种协同过滤的推荐算法,由于算法考虑到了不同语义类型资源,本文提出的算法在查准率、 F 值上均有良好的表现,证明本文提出的基于数字资源聚合的融合协同过滤推荐算法提高了推荐的准确性。

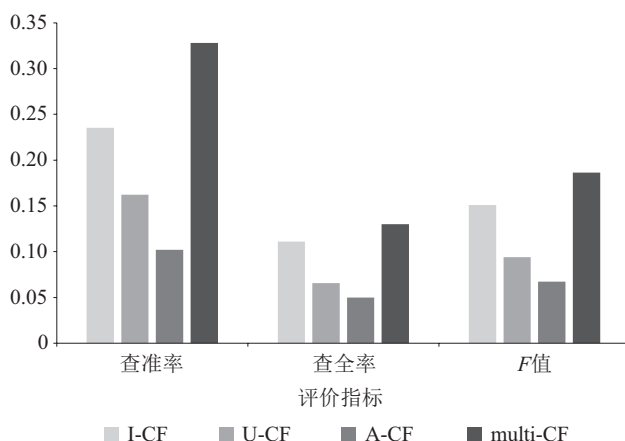


图3 4种推荐方法实验结果比较

4 结语

本文尝试将作者耦合、语义网、协同过滤推荐等方法融入图书数字资源推荐框架,提出一种基于协同过滤思想的多标签融合资源推荐方法;并以馆藏图书资源为例,通过构建的资源聚合方法计算作者关联度、用户相似度及资源相似度并获得作者关联度矩阵、用户相似度矩阵及资源相似度矩阵;采用协同过滤推荐方法,在数据集中查找相似性最高的最近邻居用户和最近邻居资源,并向用户推荐融合相似作者、相似用户及相似内容3个标签的数字资源。从实验结果可以看出,本文提出的数字资源聚合模型及资源推荐方法,既通过资源聚合挖掘数据的语义信息从而提高了相似度计算的准确性,使推荐结果具有精准化及全面化等特点,还能有效缓解数据稀疏的问题及提高推荐的新颖性,较大幅度地提高了资源推荐的精准性和资源质量。

不过本研究仍有一些不足,如实验未考虑基于作者主题的语义模型,这必然会对实验结果有所影响,由于采用协同过滤思想需要用户提供行为数据,因此尚未解决资源推荐“冷启动”的问题。后续将建立基于作者主题的LDA模型进一步挖掘作者层面的语义信息,完善语义网的概念属性及构建方法,丰富数字资源并合理选取关键词,解决“冷启动”问题,进一步提高数字资源聚合效果与推荐的质量,满足用户精准化的服务需求。

参考文献

- [1] 邱均平, 王菲菲. 基于共现与耦合的馆藏文献资源深度聚合研究探析[J]. 中国图书馆学报, 2013, 39(3): 25-33.
- [2] 朱白. 数字图书馆推荐系统协同过滤算法改进及实证分析[J]. 图书情报工作, 2017, 61(9): 130-134.
- [3] 吴晓英. 基于概率矩阵分解的馆藏数字资源智能推荐方法研究[J]. 情报理论与实践, 2014, 37(11): 94-97.
- [4] SELAMAT M H, ISA W M W, HAMID J A, et al. PTree: A tool to draw tree for Concept Relation Tree (CRT) [EB/OL]. [2018-10-20]. citeseerx.ist.psu.edu/viewdoc/download;jsessionid=038304D5120B2BCE2B6639B9C2DFACD1?doi=10.1.1.402.8265&rep=rep1&type=pdf.
- [5] 黄文碧. 基于元数据关联的馆藏资源聚合研究[J]. 情报理论与实践, 2015, 38(4): 74-79.
- [6] 严春子. 公共文化数字资源聚合服务平台建设[J]. 图书馆学研究, 2016(11): 45-47.
- [7] 胡媛, 陈琳, 艾文华. 基于知识聚合的数字图书馆社区集成推送服务组织[J]. 图书馆学研究, 2017(19): 9-17.
- [8] 毕强, 刘健. 基于领域本体的数字文献资源聚合及服务推荐方法研究[J]. 情报学报, 2017, 36(5): 452-460.
- [9] ZHAO D Z, STROTMANN A. Evolution of research activities and intellectual influences in information science 1996—2005: Introducing author bibliographic-coupling analysis[J]. Journal of the American Society for Information Science and Technology, 2008, 59(13): 2070-2086.
- [10] 刘健. 数字图书馆资源聚合与服务推荐研究[D]. 长春: 吉林大學, 2017.
- [11] 田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报(信息科学版), 2010, 28(6): 602-608.
- [12] 熊回香, 窦燕. 基于LDA主题模型的标签混合推荐研究[J]. 图书情报工作, 2018, 62(3): 104-113.
- [13] SHELTON B E, DUFFIN J, WANG Y X, et al. Linking open course wares and open education resources: creating an effective search and recommendation system[J]. Procedia Computer Science, 2010, 1(2): 2865-2870.
- [14] TSUJI K, TAKIZAWA N, SATO S, et al. Book recommendation based on library loan records and bibliographic information[J]. Procedia-Social and Behavioral Sciences, 2014, 147: 478-486.
- [15] FEDELUCIO N, PIERPAOLO B, CATALDO M, et al. Concept-based item representations for a cross-lingual content-based recommendation process[J]. Information Sciences, 2016, 374: 15-31.
- [16] 周之诚. 用户意图聚类的数字资源推荐方法[J]. 情报理论与实践, 2011, 34(6): 116-119.
- [17] 周玲元, 段隆振. 数字图书馆联盟中基于情境感知的个性化推荐服务研究[J]. 图书馆理论与实践, 2014(7): 67-69, 87.
- [18] 曾子明, 金鹏. 基于用户兴趣变化的数字图书馆知识推荐服务研究[J]. 图书馆论坛, 2016, 36(1): 94-99.
- [19] 温芳芳. 作者分类号耦合分析与作者关键词耦合分析的比较研究[J]. 情报杂志, 2017, 36(11): 186-191.
- [20] 吴彦文, 刘闯. 基于用户偏好和可疑度的推荐方法研究[J]. 计算机应用研究, 2018(12): 1-2.

作者简介

吴彦文, 女, 1971年生, 博士, 教授, 研究方向: 数字图书馆、信息资源管理。

牛晓璇, 女, 1994年生, 硕士研究生, 通信作者, 研究方向: 数字图书馆、信息资源管理, E-mail: 1467193584@qq.com。

胡炎贵, 男, 1991年生, 硕士研究生, 研究方向: 数字图书馆、信息资源管理。

王馨悦, 女, 1998年生, 本科生, 研究方向: 信息资源管理。

何秀玲, 女, 1971年生, 博士, 教授, 研究方向: 数字图书馆。

Research on Precision Recommendation Algorithm for Digital Resource with Integration Method

WU YanWen¹ NIU XiaoXuan¹ HU YanGui¹ WANG XinYue² He XiuLing³

(1. School of Physical Science and Technology, Central China Normal University, Wuhan 430079, China; 2. School of Information Management, Central China Normal University, Wuhan 430079, China; 3. National Engineering Research Center for e-Learning, Central China Normal University, Wuhan 430079, China)

Abstract: In view of the problems of information overload, heterogeneous information and unsatisfactory recommendation effect of digital resources, this paper aims to improve the traditional digital resource integration model and similarity calculation method, and combine a multi-label collaborative filtering methods to improve the accuracy of recommendation. Based on the idea of collaborative filtering recommendation, a digital resource integration method is used to calculate the similarity then find close neighbors of resources and users. Based on this, the precision of resource recommendation algorithm is constructed. Finally, the collection of library resources is taken as an example to verify the model's effectiveness. The results show that the method can effectively aggregate the digital resources, excavate the semantic information of books, and combine the user interest model to provide the users with accurate resource recommendations.

Keywords: Integration of Digital Resource; Author Coupling; Semantic Network; Collaborative Filtering

(收稿日期: 2018-11-03)