DOI:10.13530/j.cnki.jlis.2021032

钩玄提要——古籍目录智能分析工具构建

李 惠 陈 涛 侯君明 刘 丁 朱庆华 刘 炜

摘 要 古籍目录辨章学术,考镜源流,对古典学术研究具有重要的价值。本文提出古籍提要网络分析模型,用无向三部图整合古籍、人物和提要信息。在此基础上构建古籍目录智能分析工具,不仅可以自动挖掘提要中蕴藏的人物关系,与已有的古代人物知识库相关联,为知识库补充可靠而有价值的关系信息;而且综合考虑提要的元数据和正文的语义特征信息,并将其整合入推荐算法中,能为用户智能地推荐与被检索项内容、部类名、古籍名、古籍责任者相关的提要。以《四库全书总目》为实验数据集,一方面基于提要网络,从人物、古籍、提要三个层面探索不同实体间的内在联系,并就四部提要中出现的人名和古籍名开展定量研究;另一方面从作者简介、内容概述及学术评价这三种提要文本特征入手,结合元数据信息和三种常用的文献推荐算法,评估不同的语义特征对工具推荐功能准确性的影响。实验结果表明,提要文本中的内容概述及学术评价作为语义特征提炼,再结合元数据信息,效果良好,可推广应用到面向古籍的知识发现中。图 4。表 6。参考文献 51。

关键词 古籍目录 提要 网络模型 智能分析工具 数字人文

分类号 G254

Noting the Essentials: An Explorative Tool for Catalog Annotations in Chinese Rare-Book Collections

LI Hui, CHEN Tao, HOU Junming, LIU Ding, ZHU Qinghua & LIU Wei

ABSTRACT

Catalog annotations in Chinese rare-book collections, also known as Tiyaos, contain the essential information regarding a book, e.g., the author introduction, the summary, the nature and style, the version, and the critique of the corresponding book. In order to write a good Tiyao, even the most eminent scholars spent a lot of time and effort in collecting, collating, reviewing, and annotating large-scale book collections. However, confronted with large-scale rare-book collections, even though we pay huge human efforts on writing, editing, and recommending Tiyaos, this task is still time-consuming and omissions are inevitable. In this paper, we propose a Tiyao-centric network model, which integrates rare-books, historic figures, and Tiyaos into one tripartite graph. This network model is not limited to the language or scale of texts, and can be further applied to large-scale catalogues of rare-book collections.

Based on this model, we construct a Tiyao explorative tool TiyaoX. By using SPARQL, this tool can extract RDF data in related knowledge bases and provide users with information of rare-book authors or editors, for instance, occupation, imperial examination, biography, and so on. Furthermore, this tool can automatically extract the individual relations embedded in Tiyaos, enrich the existing resources with reliable and valuable relation information; in addition, this tool leverages metadata and content features of Tiyaos to recommend potential

通信作者:李惠,Email:lh9743@126.com,ORCID:0000-0001-7050-1845 (Correspondence should be addressed to LI Hui, Email:lh9743@126.com, ORCID:0000-0001-7050-1845)

interesting and query-relevant (e. g. , Tiyao content, category, book name and author relevant) Tiyaos.

In this paper, we use *Siku Quanshu Zongmu* as dataset, and on the one hand, we investigate the latent relations among rare-books, historic figures, and Tiyaos. We also make separate quantitative analysis towards person names and book names in four divisions of this dataset respectively. The results demonstrate that Tiyaos in "Ji" division contain most names of historical persons and books, and Tiyaos in "Jing" division contain names least. Tiyaos in "Zi" and "Ji" divisions have the maximum overlap of person names, and Tiyaos in "Shi" and "Zi" divisions have the maximum overlap of book names. In the constructed network, most key figures are Confucius scholars, book collectors, and bibliographers, and most important rare-books are descriptive catalogues, Confucian classics, and history books. On the other hand, we take advantage of descriptive features in Tiyao content (author introduction, content summary, and critics), and combine them with Tiyao meatada as well as three text recommendation strategies (Cosine similarity, LDA + JS distance, and Word2Vec + RWMD). Our objective is to evaluate the impacts of different content features on the accuracy of recommendation module of TiyaoX respectively. The experimental results demonstrate that the approach that integrates summary, critique information and Tiyao metadata information as content features, performs best among all results. This approach can be extended to knowledge discovery of rare-book collections which provides convenience for related professionals, scholars and enthusiasts, and improves efficiency in practice. 4 figs. 6 tabs. 51 refs.

KEY WORDS

Catalogs in Chinese rare-book collections. Tiyao. Network model. Explorative tool. Digital humanities.

0 引言

古典目录学是关于中国古代图书目录整理研究、"辨章学术、考镜源流"的一门传统学问[1],被视为"治学之门径"。提要,又称叙录、解题,是古籍目录的重要类型之一。提要是根据一定的体例编写的,对书籍的作者、卷次、内容和版本源流及其考证、评价等简要说明;提要钩玄,举其要旨,"学者览其目录,犹可想见全书之本末"[1]。

我国早在春秋时期就已出现书序一类较原始的文献提要,"《诗》《书》之序,即其萌芽"[2]。西汉末年,刘向在整理校勘皇室藏书时,"每一书已,向辄条其篇目,撮其指意,录而奏之"[3]。换言之,每书校毕,写成《叙录》一篇;在《叙录》中,罗列全书篇目于前,继之介绍作者生平、思想内容、写作价值以及学术源流等[4]。刘向《别录》之后,历代均不乏提要之作,汉有《汉书·艺文志》,唐有《隋书·经籍志》,宋有晁公武《郡斋读书志》、陈振孙《直斋书录解题》,元有马端临《文献通考·经籍考》。至清乾隆年间《四库全书总目》(以下简称《四库提要》)书成,提要更臻于大成。《四库提要》不仅"叙作者之爵里,详典籍之源流,别白是非,旁通曲证,使瑕瑜不掩,淄渑以别",而且"剖析条流、斟酌古今、辨章学术、高挹群言"[5]。

目录学家姚明达先生将传统的提要内容概括为八个方面[6]:一、著录书名和篇名;二、叙述校雠之原委,包括版本异同、篇数、书名异称、文字讹谬等;三、介绍著者生平思想;四、说明书名含义、著书原委及书的性质等;五、辨别书籍真伪;六、评论思想或史事是非;七、叙述学术源流;八、判定书的价值。古籍提要对于古籍整理研究具有重要的价值,对历代目录著作的研究也是古典文献学者关注的主要内容。在大数据和人工智能快速发展的背景下,将新技术应用于古籍整理研究,不仅可以协助读者对古籍作

出精读或泛览的选择,而且可以辅助研究者对大规模文本开展计算分析与知识发现,挖掘特定模式,提供新的思路与方法。

本文在已有古籍数字化整理成果的基础上,提出古籍提要网络分析模型,以《四库提要》为例,构建古籍目录智能分析工具,挖掘提要中蕴含的特有信息,为古典文献研究提供新的视角。本文试图解决下述问题。

(1)即便是针对小规模古籍提要数据的研究,学者们仍需要披阅大量文献,工作周期相对较长,如果从社会网络分析的角度,将提要、提要对应的人物及古籍整合入一个网络模型,研究者可以方便地探索各种实体之间的关系:不仅是人物之间的关系,还可以是提要间、古籍间的关系,甚至是跨实体之间的关系。

(2)推荐算法领域的研究体系已较成熟,但针对古代汉语从语义层面推荐文本的研究相对较少。如果能提炼出提要文本中的语义特征(即行文结构信息),并综合考虑部类名、古籍名和作者名等相关的元数据信息,继而整合入推荐算法中,可以更全面地为古籍研究学者提供参考。

(3)尽管《四库提要》在目录学领域占据重要地位,但对全部提要文本中涉及的人物和古籍,并未有过定量研究。若从数字人文的角度,对《四库提要》中的人物和古籍出现频率和共现情况开展计量或计算研究,可以为古籍专家提供新的线索和思路。

1 实践与研究回顾

传统的古籍整理完全依赖人工,不但费时费力、效率不高,而且成果的出版利用也存在种种困难;采用数字化技术对古籍进行整理和分析,可以提高分散在各地的古籍的利用效率,有效解决古籍"藏"与"用"的矛盾。

古籍数字化及自动整理研究,主要涉及古籍的自动校勘、标点、注释等方面,是数字人文研究的基础。在自动校勘领域,常娥[7]提出基于窗口匹配技术的自动校勘算法,自动发现并标记出农业古籍不同版本之间的文字差异;朱翠萍[8]针对字书,设计自动校勘程序,比对不同版本间的字条和释文字符差异。在自动标点方面,黄建年[9]利用语义和句法特征编写规则,对农业古籍进行断句和标点;胡韧奋等[10]利用深度学习模型 BERT(Bidirectional Encoder Representation from Transformers)在大规模古文文本上开展自动学习,完成句读任务。在自动注释领域,马创新等[11]计算句子相似度来实现古籍原文与注疏文献的句子对齐,并以朱熹的《论语集注》为例实现自动抽取字、词和短语三类的注释;"中国哲学书电子化计划"①将数据库中的原典文献与对应的注释文本进行"原文—注释"的链接。

古籍的数据化分析属于交叉学科的范畴,涉及古典文献学、计算机科学、语言学、历史学、文学、图书馆学、考古学等诸多学科,关联到文本挖掘、知识图谱、文本生成等多个领域的研究内容。如果说数字化是古籍内容直观的转述和再现,那么数据化就是对古籍语义的深度解析[12]。在文本挖掘领域,机器学习算法如决策树、贝叶斯网络等已运用于分析中医古籍,归纳识别有用信息,研究症状与证型的对应模式[13];秦贺然等[14]采用支持向量机的机器学习算法,统计特征词并融入实体特征,实现十本先秦典籍的自动分类;刘浏等[15]利用向量相似度和贝叶斯分类器,自动判断先秦文献的所属时代。在知识图谱方面,周莉娜等[16]设计唐诗本体模型,构建唐诗知识图谱,实现了对大规模唐诗的语义探索;陈涛等[17]将《康熙字典》中的部首、集、卷、词语间的关系,以知识图谱的形式表现出来。在文本生成领域,梁健楠等[18]利用循环神经网络学习古诗句的语义表示,根据用户输入的首句,自动生成集句诗;张开旭等[19]构建图模型结合规则,实现古文对

① https://ctext.org/introduction/zhs

联的自动应对。

随着数字人文的热潮,国内越来越多的学者采用社会网络分析的原理和方法,来展现文本内或文本间信息对象的复杂关系[20]。许超、陈小荷[21]基于《左传》和白话左传,构建词共现网络和春秋人物关系网络,研究先秦汉语词汇的整体面貌,以及春秋历史中的关键人物;李娜[22]基于《方志物产》山西分卷,抽取物产名称与其他命名实体之间的关联关系,构建社会网络,发现物产名称与别名之间的关系、物产名称与引用人物之间的关系、物产名称与用途之间的关系,以及物产在时空上的变迁研究;范文洁等[23]从《左传》描绘战争的句子中抽取进攻方和防守方来构建网络,揭示春秋时期诸侯之间的合作与对抗关系;陈蕾等[24]基于《红楼梦》中两两互动的语料,提取特征词语,定位具有权势差距的人物对,构建《红楼梦》小说中的权势网络;施晓华等[25]基于古徽州契约文书,构建交易社会网络,分析古徽州契约交易的社会结构与属性;严承希等[26]基于符号分析法,建立宋代的政治网络,阐述宋代政治网络的政治关系演化模式。但由于数据限制,以上研究成果都是以单一的年份值来定义特定年代的历史人物。

近十年来,文献编目领域从理论模型、标准规范到实践应用均发生了巨大变化[27]。随着信息技术的发展,国内很多研究机构纷纷开展古籍数字化工程,如上海图书馆的"中文古籍联合目录及循证平台"[28]收录有 1 400 余家机构的古籍馆藏目录,该平台不仅向读者提供古籍目录信息(如图像、作者、版本、馆藏地等),而且改进古籍文献检索方式,方便读者查找古籍,随时随地了解馆藏情况,实现了古籍信息的交流和资源共享。

目前,国内建设的古籍目录数据库大致可分为三类[29]:第一类是沿袭卡片目录的"分类、书名、作者"格式设立检索栏目;第二类是增加了版本项的检索,如 CALIS① 的"版本类别、出版年代、出版地点"的树状浏览;第三类是致力于将目录的所有内容,如印章、批校题跋、书影等,设立成检索栏目,如国家图书馆的古籍善本国际联合书目系统②,著录了三十余家海内外图书馆所藏古籍善本,检索项包含题跋、钤印、书影、装帧形式等 11 项。如表 1 所示,古籍目录数字化资源建设的主体主要包括公共图书馆(如国家图书馆、上海图书馆)、学术机构(如华东师范大学)、古籍数字化企业(如中华书局)等。这些单位在保护古籍的同时,满足教学、科研和市场的需求,具有很强的针对性和实用性[30],为实现古籍的数据化奠定了基础。

文献推荐在图书馆服务中已有应用,主要采用基于关联规则、基于内容、协同过滤及混合推荐的算法,重点用于现代图书和论文的检索,目前还缺乏针对古典文献的推荐研究。古典学术与现代科技相结合有着很大的必要性和可行性[31],可以拓展古典目录学公共服务的能力和方向。如果能将推荐研究和古籍提要的文本特征相结合,可以为古籍整理研究提供极大的便利。本文在现有古籍目录数据化成果的基础上,构建古籍目录智能分析工具,充分利用提要内容的语义特征,结合提要的元数据信息提出网络模型,挖掘古籍提要中所蕴涵的丰富信息以构建提要知识网络,并基于此开展知识发现服务。

2 研究方法

从数据分析的角度来看,古籍提要主要由两部分组成:提要元数据和提要内容。大部分古籍提要元数据应包括古籍题名(书名)、卷数、责任者(编撰者)、版本(可含版本年代、版本类型、版刻地、版刻者等内容)及部类(目录部类)这五个元素。

① http://rbsc.calis.edu.cn:8086/aopac/jsp/indexXyjg.jsp

② http://mylib.nlc.cn/web/guest/zhonghuagujishanbenlianheshumuxitong

表 1 具有代表性的古籍目录数字化资源列表

古籍目录数字化项目	提要信息	内容特色
中华古籍善本国际联合书目系统	未开放	书目数据及检索平台
国家珍贵古籍名录数据库①	有	书目数据及检索平台
全国古籍普查登记基本数据库②	未开放	书目数据及检索平台
中文古籍联合目录	未开放	书目数据及检索平台
大学数字图书馆国际合作计划(CADAL)③	未开放	书目数据及检索平台
华东师范大学馆藏目录④	有	书目数据及检索平台
中华古籍书目数据库⑤	有	书目数据及检索平台、提要视图
CALIS 联合目录公共检索系统	未开放	书目数据及检索平台
中文古籍联合目录及循证平台⑥	有	书目数据及检索平台、关联数据统计分析及可视化
东京大学东洋文化研究所⑦	未开放	书目数据及检索平台

2.1 构建提要网络

基于上述信息,提要的集合用 TY 指代。一篇提要 $ty \subseteq TY$ 可以表示为一个四元组集合 $ty = (P, b, ca, c)$。P 表示人物(责任者和古籍文本中出现的人物)集合,每一位人物 $p \in P$ 由 id、正式的姓名、别名以及其他描述性的属性如朝代、性别、职业、小传等组成。$b \in B$ 表示古籍(提要对应的古籍和提要文本中出现的古籍),每本古籍 $b \in B$ 由索引 id、题名、卷书、版本、版本年代及正文组成。$ca \in CA$ 指代部类,$c \in C$ 指代提要内容,即字序列 $\{w_1, w_2, \cdots, w_n\}$。我们将提要网络模型定义为无向三部图 $G = (V, E)$,其中节点集合 $V \subseteq P \cup TY \cup B$ 代表人物 P、提要 TY 和古籍 B 的集合,边集合 $E \subseteq V \times V \times V$ 代表节点之间的关系。使用 aw、tw 和 mw 作为人物节点和提要节点之间边的属性,区分古籍的著者、编者和提要正文涉及的人物;用 bw 和 cw 作为古籍名节点和提要节点之间边的属性,区分提要对应的古籍名和提要正文涉及的古籍名。

提要网络可以帮助我们抽取和分析提要、古籍和责任者之间的关系,相比于三部图里混合的多种关系,本文用特定的子图信息来分析特定的关系,将提要网络模型投射为人物关联网络、古籍关联网络和提要关联网络,更加准确清晰,如图 1 所示。本文抽取人名和古籍名在提要正文中的共现信息,旨在突破古籍语料获得渠道的限制,挖掘提要中隐含的学术流派和发展脉络,探寻古籍浩繁卷帙之间的潜在联系。

(1)古籍关联网络。基于上文定义的提要网络模型,古籍关联网络可以定义为无向图 $G_b = (V_b, E_b)$。节点集合 $V_b \subseteq B$ 代表古籍,而有向边集合 $E_b \subseteq V_b \times V_b$ 代表内容上的关联或在提要正文中的共现关系。如果两本古籍文本的相似度超过一定阈值或在同一篇提要中均有涉及,那么它们在古籍关联网络中即存在连接关系。本文侧重于分析提要正文中古籍的共现关系。

① http://192.168.42.10/nlcab/public! mlSearch.action
② http://202.96.31.78/xlsworkbench/publish
③ http://www.cadal.zju.edu.cn/Index.action
④ http://202.120.82.40/*chx
⑤ http://bib.ancientbooks.cn/docGuji
⑥ https://gj.library.sh.cn/index
⑦ http://www3.ioc.u-tokyo.ac.jp/kandb.html

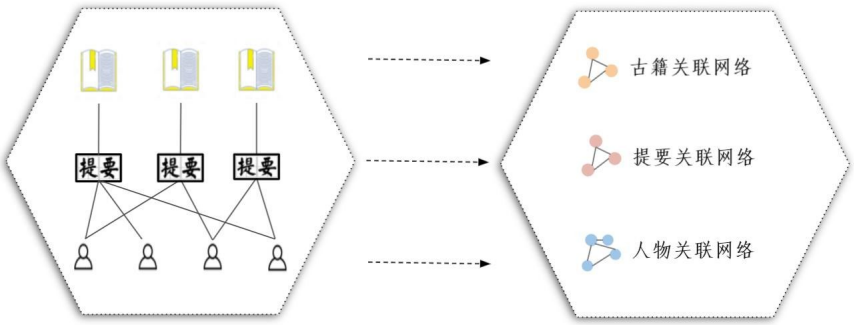


图 1　提要网络模型及投射的三种子图示例

(2)人物关联网络。人物关联网络可以定义为无向图 $G_p=(V_p,E_p)$,节点集合 $V_p \subseteq P$ 代表人物,而无向边集合 $E_p \subseteq V_p \times V_p$ 代表他们的合作编撰关系或在古籍提要正文中的共现关系。如果两人联合编撰同一本古籍或在同一篇提要中均有涉及,那么他们在人物关联网络中就是相连接的。本文将这两种关系类型作为属性附加给每条边,在后续实验中会基于边的属性,将人物关联网络投射为人物共现网络和责任者合作网络,以便分析更加直观和便捷。

(3)提要关联网络。若两篇提要在内容上存在一定的相似性,那么它们在提要关联网络中就是连接的。该网络可以定义为 $G_{ty}=(V_{ty},E_{ty})$,节点集合 $V_{ty} \subseteq TY$ 代表提要,而有向边集合 $E_{ty} \subseteq V_{ty} \times V_{ty}$ 代表提要间的内容关联。本文将提要之间的这种关联性具体计算为:

$$sim(ty_1,ty_2)=\alpha_1 \times sim(p_1,p_2)+\alpha_2 \times sim(b_1,b_2)+\alpha_3 \times sim(ty_1,ty_2)+\alpha_4 \times sim(ca_1,ca_2)$$

其中 $\alpha_1+\alpha_2+\alpha_3+\alpha_4=1$。$sim(p_1,p_2)$,$sim(b_1,b_2)$,$sim(ty_1,ty_2)$ 和 $sim(ca_1,ca_2)$ 分别计算责任者、题名、提要文本内容以及部类之间的相似性。关于责任者姓名、古籍题名和部类名的相似度计算,采用字符串完全匹配的算法,如果完全相同,则相似度为 1,其他情况皆为 0。

$$sim(p_1,p_2)=\begin{cases}1, & p_1=p_2\\ 0, & p_1\neq p_2\end{cases}$$

$$sim(b_1,b_2)=\begin{cases}1, & b_1=b_2\\ 0, & b_1\neq b_2\end{cases}$$

$$sim(ca_1,ca_2)=\begin{cases}1, & ca_1=ca_2\\ 0, & ca_1\neq ca_2\end{cases}$$

本文共选取三种方法:余弦相似度、话题模型 LDA 和 JS 距离、词向量模型和 RWMD 词移距离,分别计算提要文本之间的相似度,并将结果交于古籍专家检验,比较各种方法在提要相似性计算中的有效性。

(1)余弦相似度(Cosine Similarity)。余弦相似度是计算文本相似度的常用算法[32],通过将文本映射到向量空间,计算两个向量的夹角余弦值,来衡量文本之间的差异性。

(2)话题模型 LDA(Latent Dirichlet Allocation)和 JS 距离(Jensen-Shannon Distance)。LDA[33]是融合了贝叶斯统计和优化算法的生成式概率模型,该模型将每篇文本表示为话题的混合分布,而每个话题是词语的概率分布[34]。该算法可有效降低文本表示维度,提升后续相似度计算的效率[35],常用于抽取文本话题。JS 距离计算两种概率分布之间的差异性,具有对称性,比 KL 距离更适合衡量语义关联的主题[36]。

(3)词向量模型(Word2vec)和 RWMD 词移距离(Relaxed Word Mover Distance)。近年来,深度学习成为学术界和工业界的热点[37]。

Word2vec[38]作为深度学习的语言建模算法,受到自然语言处理领域的高度重视。Word2vec 将概率模型与人工神经网络相结合,利用文本中词语的上下文信息,生成每个词对应的词向量[39]。WMD 距离源于 EMD (Earth Mover Distance) 模型[40],通过计算一个文本的词向量"移动"为另一文本的词向量所需的最小距离,来度量文本之间的差异。RWMD 在 WMD 的基础上,放宽了约束条件,降低了计算复杂度,提升了计算效率。

本文分别采用特征向量中心性(Eigenvector Centrality)和 Louvain 算法挖掘网络中的重要节点和潜在社群。一个节点的重要程度不仅与自身有关,而且与其相邻节点的重要程度相关[41]。相邻节点越多,这个节点就越重要。特征向量中心性一般计算为对应图的邻接矩阵的最大特征向量值[42]。Louvain 算法基于模块度(Modularity)的相关概念,通过两步迭代,即模块度优化和社群聚合,挖掘网络中的社群结构[43],该算法效率高,准确率也有保证[44]。

一篇古籍提要的文本大致可划分为三节:位于开头的作者简介、中间部分的内容概述(含著述体例、版本源流等)、篇尾的学术评价。我们邀请古籍领域的学者将实验数据中的每篇提要划分为三部分并加以标注,在实验阶段计算分析不同的提要文本特征和对相似性计算可能产生的影响。

2.2 古籍目录智能分析工具

基于上述提要网络模型,本文构建古籍目录智能分析工具(具体分析流程见图 2),侧重于基于提要网络的人物关系提取和基于提要文本语义特征的智能推荐,这里简述工具所具有的四种特色模块功能。

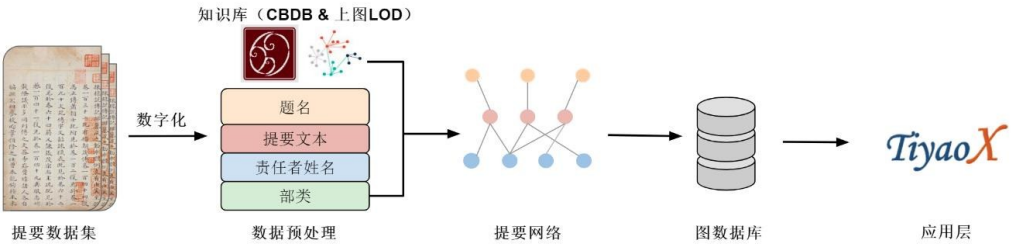


图 2 古籍目录智能分析流程

(1)古籍责任者信息。基于用户选择的古籍名称和版本信息,本工具使用 SPARQL 语句查询不同古籍的责任者在中国历代人物传记资料库 (CBDB)① 和上海图书馆人名规范库 (NAD)②的 RDF 信息,向用户提供责任者的职业、科举、传记、相关作品等信息,以及库内存在的责任者的社会关系。

(2)抽取编撰关系并可视化。本工具基于提要网络模型,抽取古代学者之间的编撰关系、古籍书名在提要中的共现关系,生成可视化的人际网络;同时将抽取的编撰关系整合入已有的关联数据中。本工具不单是利用外部资源,而且也为已有资源扩充新的信息。在下文的实验中,将提要网络中抽取的关系与外部知识库已有的关系进行信息比对,从而说明网络分析功能的有效性。

(3)比较不同版本之间的提要异同。古籍常有多种版本,本工具可自动显示其他版本的

① https://projects.iq.harvard.edu/cbdb/home.
② http://names.library.sh.cn.

提要基本信息,并且比较古籍不同版本对应的提要文本之间的相似性。

(4)向读者推荐感兴趣的提要。考虑到读者可能对相似主题的提要感兴趣,本工具提取提要的语义特征,并结合元数据信息,计算提要文本之间的相似度,智能地为读者推荐相似度较高的古籍提要信息。本文会在后续实验中,结合不同的相似性计算方法,详细评估不同的提要文本特征对推荐功能准确性的影响。

3 实验

本研究选择《四库提要》文本作为实验数据。《四库提要》对《四库全书》所收每本书的作者生平、主要内容、著述体例、版本源流等均有涉及,且考订文字异同讹误、评议得失,自成书以来一直受到学者的高度重视[45]。实验选取上海图书馆馆藏电子资源《四库提要》文本,共 10 297 条,该文本已经过重新整理,正文中涉及书名均使用书名号。通过词性标注工具 jiebaR 和古籍责任者姓名词表来提取《四库提要》正文中的人名,用正则表达式提取正文中标有书名号的书名,初步共抽取人名 10 424 个,书名 4 536 个。考虑到一位古人可能会有多个名号指代,比如孔子,又称为仲尼、圣人、宣父、孔圣人、至圣先师等,我们请相关领域的专家进行人工校验,规范每个人的名字,将一人多名标准化为单人单名,将高频的姓名均通过史料考证,避免多人同名的歧义。同时我们也发现,提要文本中有书名缩略的情况,可能会引起歧义,如《序》《传》《本义》等,也请专家人工校验,并将高频出现的书名做了标准化处理。

标准化处理之后,共得到人名 7 214 个,书名 4 148 个(四部中的人名和书名已考虑重叠情况,去重计算)。集部提要涉及人物和古籍最多,共出现 3 831 个人名和 2 181 个古籍名,经部提要涉及古籍和人物最少,共出现 1 892 个人名和 1 002 个古籍名,史部出现 2 953 个人名和 1 392 个古籍名,子部出现 3 211 个人名和 1 719 个古籍名。分别统计每两部提要间的人物和古籍的共现情况(见表 2),并用 Jaccard 系数计算重合度。相比其他部类,子部和集部提要的人物重合比例最高(29.7%),史部和子部提要的古籍重合比例最高(23.7%),经部和集部提要无论是人物还是古籍的重合比例都是最小的;子部和集部高频出现的人物重合较多,如苏轼、朱熹、王守仁等,这是因为他们不仅是思想家也是文学家,所以两部提要中均有提及;史部和子部提要中高频出现的古籍名较多重合,如《直斋书录解题》《文献通考》《永乐大典》等,这与提要的性质有关,这些学术价值极高的藏书目录、类书、政书等是编纂《四库全书》和编写《四库提要》的重要参考资料。

表 2 《四库提要》每两部间人物和古籍的共现情况

《四库提要》	人物(%)	古籍(%)
经部史部	20.7	19.7
经部子部	23.7	21.9
经部集部	20.0	16.4
史部子部	29.2	23.7
史部集部	28.7	20.1
子部集部	29.7	23.2
四部均出现	7.4	6.0

(1)人物共现网络。实验基于《四库提要》语料构建了人物关联网络,并将其投射为人物共现网络和责任者合作网络。人物共现网络共包含 6 196 个节点和 92 862 条边。根据上文提到的网络度量方法,计算了网络中的重要人物和语料中的高频人物并排序(如表 3 前两列所示)。网络中排名靠前的多是著名的儒学家如孔子、孟子等,其次是藏书家和目录学家,如陈振孙、朱彝尊(考虑到朱所著《经义考》的重要性,在此将其归为藏书家和目录学家之列)等,接着是文学家,如苏轼等。考虑到清代帝王在《四库全书》和《四库提要》编纂中的特殊地位,在试验中略去与之相关的重要性度量和社群划分。通过上文提到的 Louvain 算法,共检测到人物共现网络 55 个社群(见图 3),其中只有 5 个社群的节点数超过 500,约 76.3%的社群内部节点数不超过 5 个。图 3(b)网络中第一大社群共包含 1 571 个节点,多数是明代文学家如刘基、宋濂、高启等;第二大社群共包含 1 211 个节点,多为著名的目录学家和藏书家,如晁公武、王应麟、焦竑等。需要说明的,表 3 中从《四库提要》文本中人物出现频率角度计算的高频学者,大部分出现在网络中第三和第四大社群,并非最大社群。

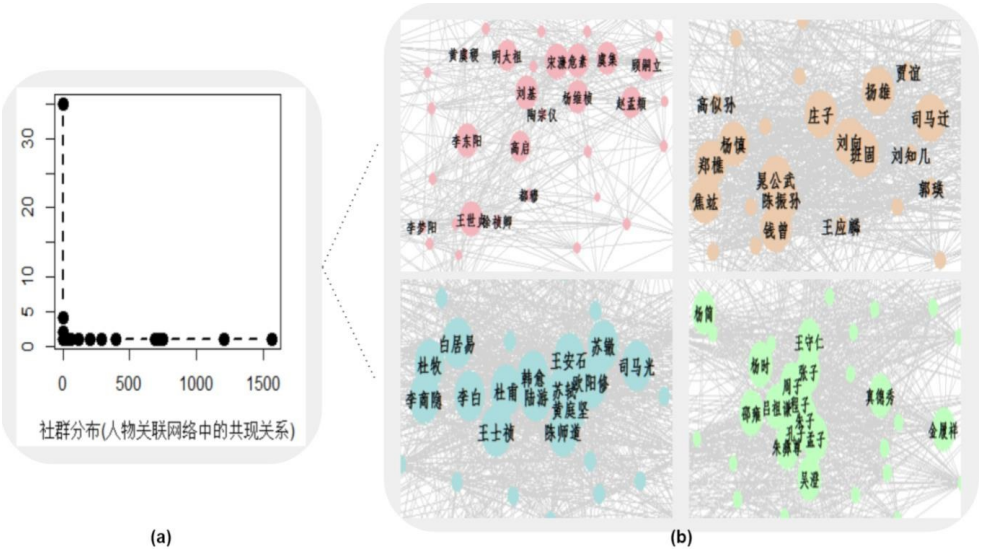
表 3 人物共现网络中的重要学者、责任者合作网络中的重要责任者、《四库提要》中的高频学者及《四库全书》中的高频责任者

重要学者	高频学者	重要责任者	高频责任者
朱熹	朱熹	刘基	毛奇龄
程子①	苏轼	宋濂	杨慎
孔子	朱彝尊	赵埙	王士祯
孟子	陈振孙	瞿庄	陈继儒
朱彝尊	欧阳修	乐韶凤	朱熹
陈振孙	王守仁	王僎	顾炎武
吕祖谦	孔子	朱右	陆深
苏轼	王士祯	朱廉	魏裔介
邵雍	孟子	邹孟达	孙承泽
周子	晁公武	汪广洋	吕祖谦

(2)责任者合作网络。将《四库提要》中记录的《四库全书》各书责任者分别做了统计,构建了责任者合作网络,共由 574 个节点(人物)和 663 条边(编撰关系)组成。将网络中存在的编撰者合作关系与 CBDB 人物传记库中可查询到的人物关系信息对照分析,如表 4 所示,CBDB 中共查询到 502 人和 6 934 条关系信息,但只有 28 条与本网络中的人物关系有交集。因此可以认为,基于古籍提要元数据信息构建的责任者合作网络,可以为已有的古代人物知识库补充可靠而有价值的关系信息。

本文同样计算了网络中的重要责任者和语料中的高频责任者并排序(如表 3 后两列所示),可以发现,中介中心性的得分较高者与语

① 《四库提要》中北宋理学家程颢和程颐的名字总是以“程子”“二程子”“二程”的形式出现,在预处理中,统一转换成“程子”。



注:图(a) x 轴代表人物共现网络中社群中的人数,y 轴代表该具体数值在所有社群数值中的频次;图 (b)由左至右由上至下选取了共现网络中最大的四个人物社群,并将网络中的重要人物标示出来。

图 3 《四库提要》人物共现网络

表 4 CBDB 数据库中可查询到的提要相关责任者和人物关系

	人物数量	人物关系数量	重合数量
CBDB	502	6 934	28
责任者合作网络	574	663	

料中的高频责任者列表有一定出入。语料中的高频责任者多是中国历史上著名的高产学者,比如陈继儒、顾炎武、毛奇龄等,而网络中的重要人物乐韶凤、宋濂、刘基等在网络中都与十个以上学者有连接,因为他们同时兼任官方的丛书、类书主编或主管[46],是丛书、类书编撰团队中的核心成员。这是因为网络中关注的是责任者之间的互动关系,即就同一本书,不同的编撰者之间历时的合作频繁程度。换言之,网络中的人并不全是同代人,但他们都编辑注解过同一部古籍。比如《阴符经解》有七个责任者,黄帝、姜太公、范蠡、鬼谷子、张良、诸葛亮和李筌,他们之间这种历时的编撰合作关系,可以帮助研究者发现不同年代、不同背景的历史人物之间的兴趣交集。

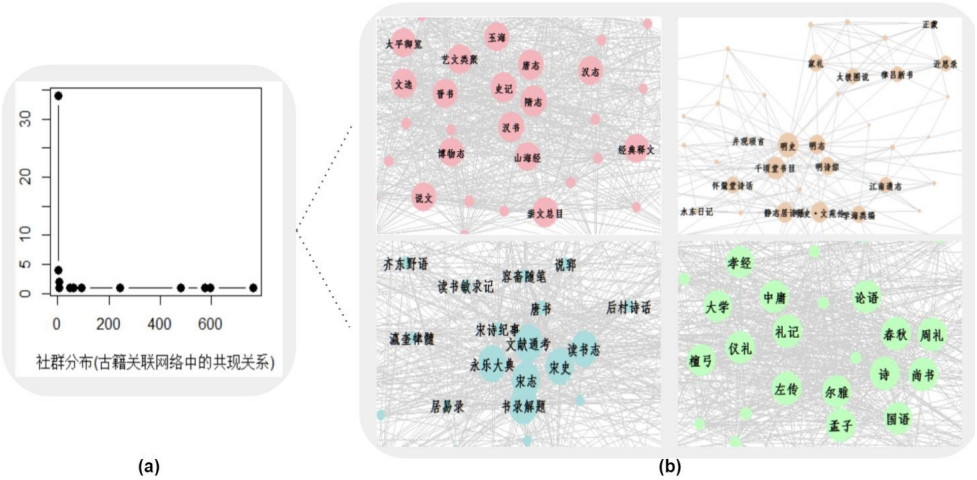
(3)古籍共现网络。实验也构建了古籍共现网络,该网络由 2 957 个节点和 29 815 条边组成。同样计算了网络中的重要古籍和语料中的高频古籍并排序(见表 5)。表中排名前十的古籍大致可分为三类,第一类是叙录式目录,如晁公武的《郡斋读书志》、陈振孙的《直斋书录解题》等,这些均为宋代私人藏书目录,学术水平很高,受到历代学者的重视和称赞,并对后世目录学的发展起到重要的引导作用①,是《四库提要》参考征引的重要资料[47];第二类是儒家典籍,如《周易》《春秋》等,符合清朝统治者对儒学的推崇;第三类为史书,如《宋史》《明史》等。

① https://gj.library.sh.cn/ancientBookCatalogue/list

表 5　古籍共现网络中的重要古籍及《四库提要》中的高频古籍

重要古籍	高频古籍
《永乐大典》	《直斋书录解题》
《直斋书录解题》	《永乐大典》
《宋史》	《周易》
《宋史·艺文志》	《明史》
《郡斋读书志》	《宋史》
《周易》	《春秋》
《文献通考》	《宋史·艺文志》
《经义考》	《经义考》
《宋史·艺文志》	《左传》
《春秋》	《郡斋读书志》

同样运用 Louvain 算法,在古籍共现网络中共检测出 49 个社群(见图 4),网络中的社群分布也呈现出明显的规律性,只有 3 个社群的节点数超过 500,约 83.7%的社群内部节点数不超过 5 个。表 5 中出现的高频古籍大部分出现在图 4(b)网络中第三和第四大的社群,而网络中第一大社群共包含 766 个节点,多数为史书如《史记》《汉书》等和史书包含的目录书如《汉书·艺文志》《隋书·经籍志》等;第三大社群共包含 597 个节点,包含私人书目如《郡斋读书志》《直斋书录解题》等、类书如《永乐大典》、诗歌资料汇集如《宋诗记事》等。



注:图(a)x 轴代表古籍共现网络中社群中的古籍数,y 轴代表该具体数值在所有社群数值中的频次;图(b)由左至右由上至下选取了共现网络中最大的四个社群,并将网络中的重要古籍名标示出来。

图 4　古籍共现网络

(4)融入文本语义特征的提要推荐。首先从实验数据集中随机选取100条提要,作为测试数据,采用2.1中所设计的相似性度量函数计算提要之间的相似性。测试集中的每一条提要,都会计算它与实验数据集中所有提要的相似性,并根据结果,推荐排名最高的5个提要。同时请古籍专家来评测推荐结果的正确性,并用Kappa统计量[48]度量专家判断的一致性。

基于第一步内容相似性筛选的提要数据,进一步度量提要文本中不同语义特征对相似性计算的影响。选取了六种文本的组合方式,即提要全文、内容概述(含著述体例、版本源流等)、学术评价、内容概述与学术评价的组合,以及作者简介分别与内容概述、学术评价的组合,采用2.1中介绍的三种方法,即余弦相似度、话题模型LDA+JS距离、Word2vec+RWMD距离开展计算,并请专家标注正确性。此时,只考虑提要文本,不加入元数据信息,即$\alpha_1=\alpha_2=\alpha_4=0$,$\alpha_3=1$。综合选取三种方法,即最大似然估计[49]、话题之间的距离计算[50]和话题之间的连贯性[51]的方法,确定本次实验测试集中的提要话题数为67个。不拘泥于古籍文本分词限制,采用N元语法(n-gram),结合中文分词工具jiebaR对提要开展一到二元字词组抽取工作,并将Word2vec的上下文窗口长度设为3。每组实验中,专家们的Kappa统计量均在70%以上,证明他们的标注具有较好的一致性。本文由此计算推荐的准确性,即专家判断正确的提要总数占所有推荐结果的百分比。取专家评测的准确性均值作为最终结果,如表6所示。

表6 结合题名、责任者姓名、部类名、提要内容的综合相似性计算的准确性评测

$\alpha_1=\alpha_2=\alpha_4=0$, $\alpha_3=1$			
	余弦相似度(%)	LDA+JS(%)	Word2vec+RWMD(%)
全文	58	47.2	38.6
内容概述	55.8	38.4	26.2
学术评价	28.8	12.8	19.2
作者简介+内容概述	46.6	44.2	28.4
作者简介+学术评价	30.8	49.6	27.8
内容概述+学术评价	62	47	26
余弦相似度			
	$\alpha_1=\alpha_2=\alpha_3=\alpha_4=0.25$(%) 平衡性		$\alpha_1=\alpha_2=\alpha_4=0.1,\alpha_3=0.7$(%) 提要中心性
内容概述+学术评价	76		75
全文	69		71

从表6可以看出,LDA和Word2vec的方法针对本研究的文本并没有达到预期的效果,这可能和提要文本简明扼要的属性有关;余弦相似度和文本特征"内容概述及学术评价"这两部分的组合,效果最好,且用时最短。于是我们选取这种组合,加入元数据信息责任者姓名、古籍名以及部类名进行相似性计算,并设置两种权重。一种是平衡性,均衡考虑元数据信息和文本特征,$\alpha_1=\alpha_2=\alpha_3=\alpha_4=0.25$;另一种是侧重提要文本内容,$\alpha_1=\alpha_2=\alpha_4=0.1$,$\alpha_3=0.7$。度量结果如表6后三行所示。最后发现,文本特征"内容概述及学术评价"加上平衡性参数,即简约低成本的度量方法加上特定的语义特征,可以取得良好的推

荐效果。

基于上文的计算结果,本文将提要相似性的阈值设为 0.25,构建提要关联网络,网络中共有 4 038 个节点和 612 844 条边。根据特征向量中心性的计算结果,网络中心提要所对应的古籍为《四库全书》史部的《元史》。除《三经附义》在经部之外,《明史》《资治通鉴后编》和《金史》也均在史部。《四库全书》由经、史、子、集四个部分组成,史部包含了很多珍贵的史书,史学价值很高,因而所对应的古籍提要在网络中处于中心位置,合情合理。我们同时也注意到,网络中有一些孤立的节点,和其他节点没有连接,即和大部分提要的内容相似性很低,这些提要对应的古籍属于集部,多为个人编撰的词曲集,比较"小众"。和经、史、子三部不同,集部更侧重戏剧、音乐和文学作品,各具风格特色,因而在网络中相对更容易"边缘化"。

4 结论

将智能计算和语义分析技术恰当地应用于古籍整理领域,既可以提高工作效率和古籍的利用率,为古籍整理提供捷径,又有助于古籍内容的深度挖掘。基于此,本文提出古籍目录智能工具,侧重于提要所关联的信息和提要内容中蕴含的关系。本工具基于构建的提要网络模型,投射出三种子网络,不仅可以抽取古籍和人物在大规模提要文本中的共现关系,而且可以挖掘古代学者的历时编撰合作关系,为数字人文领域已有的古代人物知识库补充可靠而有价值的关系信息,是对古籍提要知识发现的一次有力尝试。本工具不局限于提要数据集的规模,可辅助古籍领域的文献查阅、辞书编撰、主题查找、统计分析等层面的研究。本文在《四库提要》数据集上开展试验,结合提要的语义特征和文本相似度的计算方法,对相关提要的推荐准确性展开度量,发现融入特定的文本语义特征可以显著提高提要推荐的准确率。本文针对相关提要的推荐评估所采用的实验方法、特征选取、实验步骤及结果,可以为从事相关研究的学者提供借鉴和参考。

下一步的工作计划从两方面开展:①进一步完善和改进工具,将自然语言处理的技术如命名实体抽取、关键信息抽取等,引入工具,扩充系统的整体功能,为古籍目录的自动编撰提供新的研究视角。②进一步扩展研究领域,将古籍目录智能工具应用于近代文献领域如近代报刊研究等,辅助相关研究人员高效准确地获取文献信息编写摘要,提高对文献的利用率。

参考文献

[1] 王圻. 续文献通考:卷一六八[M]. 上海:上海古籍出版社,2002. (Wang Qi. Xuwenxiantongkao: volume 168 [M]. Shanghai: Shanghai Classics Publishing House, 2002.)

[2] 余嘉锡. 目录学发微[M]. 北京:商务印书馆,2011:7. (Yu Jiaxi. Muluxuefawei[M]. Beijing: Commercial Press, 2011:7.)

[3] 班固. 汉书·艺文志[M]. 上海:上海古籍出版社,1986:527. (Ban Gu. Hanshu· yiwenzhi [M]. Shanghai: Shanghai Classics Publishing House, 1986:527.)

[4] 张舜徽. 中国校雠学叙论[J]. 华中师院学报(哲学社会科学版),1979(1):64-78. (Zhang Shunhui. A discussion on Chinese collation[J]. Journal of Central China Normal University (Humanities and Social Sciences), 1979(1):64-78.)

[5] 余嘉锡. 四库提要辨证[M]. 北京:中华书局,1980:48. (Yu Jiaxi. A discussion on *Siku Tiyao*[M]. Beijing: Zhonghua Book Company, 1980:48.)

[6] 姚明达. 中国目录学史[M]. 北京:商务印书馆,2014:36-40. (Yao Mingda. History of China catalog[M]. Beijing: The Commercial Press, 2014:36-40.)

[7] 常娥. 古籍智能处理技术研究——农业古籍自动编纂和自动校勘的研究[D]. 南京:南京农业大学,2007. (Chang E. Research on intelligent processing technology of ancient book: automatic compilation and automatic version comparison and analysis of agricultural ancient book [D]. Nanjing: Nanjing Agricultural University,2007.)

[8] 朱翠萍. 古籍字书计算机辅助校勘初探[J]. 渤海大学学报(哲学社会科学版),2018(6):105-109. (Zhu Cuiping. Computer-aided emendation on rare-book dictionaries[J]. Journal of Bohai University (Philosophy & Social Science Edition),2018(6):105-109.)

[9] 黄建年. 农业古籍的计算机断句标点与分词标引研究[D]. 南京:南京农业大学,2009. (Huang Jiannian. Research on automation of sentence segmentation, punctuation and word segmentation of agricultural ancient books[D]. Nanjing: Nanjing Agricultural University,2009.)

[10] 胡韧奋,李绅,诸雨辰. 基于深层语言模型的古汉语知识表示及自动断句研究[J]. 中文信息学报,2021(4):8-15. (Hu Renfen, Li Shen, Zhu Yuchen. Knowledge representation and sentence segmentation of ancient chinese based on deep language models[J]. Journal of Chinese Information Processing,2021(4):8-15).

[11] 马创新,陈小荷,曲维光. 注疏文献中的注释语句自动分析[J]. 计算机科学,2012(10):220-223. (Ma Chuangxin, Chen Xiaohe, Qu Weiguang. Automatic analysis of comments in commentary literature[J]. Computer Science,2012(10):220-223.)

[12] 李明杰. 数字环境下古籍整理范式的传承与拓新[J]. 中国图书馆学报,2015(5):99-110. (Li Mingjie. The inheritance and innovation of ancient book collation paradigm in the digital environment[J]. Journal of Library Science in China,2015(5):99-110.)

[13] 马梦羽,沈璐,文天才,等. 数据挖掘技术在中医诊疗数据分析中的应用[J]. 中国中医药信息杂志,2016(7):132-136. (Ma Mengyu, Shen Lu, Wen Tiancai, et al. Application of data mining technology for data analysis of TCM diagnosis and treatment[J]. Chinese Journal of Information on Traditional Chinese Medicine,2016(7):132-136.)

[14] 秦贺然,刘浏,李斌,等. 融入实体特征的典籍自动分类研究[J]. 数据分析与知识发现,2019(9):68-76. (Qin Heran, Liu Liu, Li Bin, et al. Automatic classification of ancient classics with entity features[J]. Data Analysis and Knowledge Discovery,2019(9):68-76.)

[15] 刘浏,李斌,曲维光,等. 先秦词汇的时代特征自动获取及文献时代的自动判定[J]. 中文信息学报,2013(5):107-113. (Liu Liu, Li Bin, Qu Weiguang, et al. The automatic acquisition of Pre-Qin word's property of times and the automatic classification of document's times[J]. Journal of Chinese Information Processing,2013(5):107-113.)

[16] 周莉娜,洪亮,高子阳. 唐诗知识图谱的构建及其智能知识服务设计[J]. 图书情报工作,2019(2):24-33. (Zhou Lina, Hong Liang, Gao Ziyang. Construction of knowledge graph of Chinese Tang poetry and design of intelligent knowledge services[J]. Library and Information Science,2019(2):24-33.)

[17] 陈涛,杨开漠.《康熙字典》的古汉语知识图谱构建方法研究[J]. 五邑大学学报(自然科学版),2019(4):58-64. (Chen Tao, Yang Kaimo. Research based on the *Kangxi Dictionary* on the construction approach for an ancient Chinese language knowledge graph[J]. Journal of Wuyi University (Natural Science Edition),2019(4):58-64.)

[18] 梁健楠,孙茂松,矣晓沅,等. 基于神经网络的集句诗自动生成[J]. 中文信息学报,2019(3):126-135. (Liang Jiannan, Sun Maosong, Yi Xiaoyuan, et al. Neural network-based jiju poetry generation[J]. Journal of Chinese Information Processing,2019(3):126-135.)

[19] 张开旭,孙茂松. 统计与规则相结合的古文对联应对模型[J]. 中文信息学报,2009(1):100-105. (Zhang Kaixu, Sun Maosong. A Chinese couplet generation model based on statistics and rules[J]. Journal of Chinese Information Processing,2009(1):100-105.)

[20] 王军. 从人文计算到可视化——数字人文的发展脉络梳理[J]. 文艺理论与批评,2020(2):18-23. (Wang Jun. From humanities computing to visualization: the evolution of digital humanities[J]. Theory and Criticism of

Literature and Art, 2020(2): 18-23.)

[21] 许超,陈小荷.《左传》中的春秋社会网络分析[J]. 南京师范大学文学院学报,2014(1):179-184. (Xu Chao, Chen Xiaohe. Social network analysis of Spring and Autumn period based on *Zuozhuan*[J]. Journal of School of Chinese Language and Culture Nanjing Normal University, 2014(1): 179-184.)

[22] 李娜,包平. 方志类古籍中物产名与别名关系的可视化——基于社会网络分析技术视角[J]. 图书馆论坛,2017(12):108-114. (Li Na, Bao Ping. Visual exploration of the relationship between produce names and their alias in ancient local chronicles[J]. Library Tribune, 2017(12): 108-114.)

[23] 范文洁,李忠凯,黄水清. 基于社会网络分析的《左传》战争计量及可视化研究[J]. 图书情报工作,2020(6):90-99. (Fan Wenjie, Li Zhongkai, Huang Shuiqing. A study on the visualization and metric analysis of war in *Zuozhuan* based on social network analysis[J]. Library and Information Service, 2020(6): 90-99.)

[24] 陈蕾,胡亦旻,艾苇,等.《红楼梦》中社会权势关系的提取及网络构建[J]. 中文信息学报,2015(5):185-193,203. (Chen Lei, Hu Yimin, Ai Wei, et al. Extraction of power relationship and its corresponding social network in the *Story of Stone*[J]. Journal of Chinese Information Processing, 2015(5): 185-193, 203.)

[25] 施晓华,王昕. 数字人文社会网络分析方法应用与研究[J]. 图书馆杂志,2020(5):93-99. (Shi Xiaohua, Wang Xin. Research and application of digital humanities with social network analysis[J]. Library Journal, 2020(5): 93-99.)

[26] 严承希,王军. 数字人文视角:基于符号分析法的宋代政治网络可视化研究[J]. 中国图书馆学报,2018(5):87-103. (Yan Chengxi, Wang Jun. Digital humanistic perspective: a study on the visualization of political network in Song dynasty based on symbolic analysis[J]. Journal of Library Science in China, 2018(5): 87-103.)

[27] 胡小菁. 文献编目:从数字化到数据化[J]. 中国图书馆学报,2019(3):49-61. (Hu Xiaojing. Cataloging from digitization to datafication[J]. Journal of Library Science in China, 2019(3): 49-61.)

[28] 夏翠娟,林海青,刘炜. 面向循证实践的中文古籍数据模型研究与设计[J]. 中国图书馆学报,2017(6):16-34. (Xia Cuijuan, Lin Haiqing, Liu Wei. Designing a data model of Chinese ancient books for evidence-based practice[J]. Journal of Library Science in China, 2017(6): 16-34.)

[29] 童正伦. 古籍书目数据库析评[J]. 图书馆理论与实践,2015(12):100-106. (Tong Zhenglun. Review on Chinese historical bibliographic database[J]. Library Theory and Practice, 2015(12): 100-106.)

[30] 张惠婷. 网上中文古籍数字化资源建设现状分析[D]. 沈阳:辽宁大学,2017. (Zhang Huiting. Analysis on the construction of digital resources of Chinese ancient books on the Internet[D]. Shenyang: Liaoning University, 2017.)

[31] 柯平,刘旭青. 中国目录学七十年:发展回溯与评析[J]. 中国图书馆学报,2019(5):101-111. (Ke Ping, Liu Xuqing. 70 Years of Chinese bibliography development: retrospect and analysis[J]. Journal of Library Science in China, 2019(5): 101-111.)

[32] 张振亚,王进,程红梅,等. 基于余弦相似度的文本空间索引方法研究[J]. 计算机科学,2005(9):162-165. (Zhang Zhenya, Wang Jin, Cheng Hongmei, et al. An approach of spatial index of text information based on cosine similarity[J]. Computer Science, 2005(9): 162-165.)

[33] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022.

[34] 张志飞,苗夺谦,高灿. 基于 LDA 主题模型的短文本分类方法[J]. 计算机应用,2013(6):1587-1590. (Zhang Zhifei, Miao Duoqian, Gao Can. Short text classification using latent dirichlet allocation[J]. Journal of Computer Applications, 2013(6): 1587-1590.)

[35] 孙昌年,郑诚,夏青松. 基于 LDA 的中文文本相似度计算[J]. 计算机技术与发展,2013(1):217-220. (Sun Changnian, Zheng Cheng, Xia Qingsong. Chinese text similarity computing based on LDA[J]. Computer Technology and Development, 2013(1): 217-220.)

[36] 唐晓波,祝黎,谢力. 基于主题的微博二级好友推荐模型研究[J]. 图书情报工作,2014(9):105-113.

(Tang Xiaobo, Zhu Li, Xie Li. Two-level microblog friend recommendation based on topic model[J]. Library and Information Service, 2014(9): 105-113.)

[37] 杨河彬,贺樑,杨静. 一种融入用户点击模型 Word2Vec 查询词聚类[J]. 小型微型计算机系统,2016,37(4):38-43. (Yang Hebin, He Liang, Yang Jing. Query clustering using CT-Word2Vec model[J]. Journal of Chinese Computer Systems, 2016, 37(4): 38-43.)

[38] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems, 2013: 3111-3119.

[39] 李跃鹏,金翠,及俊川. 基于 word2vec 的关键词提取算法[J]. 科研信息化技术与应用,2015(4):54-59. (Li Yuepeng, Jin Cui, Ji Junchuan. A keyword extraction algorithm based on Word2vec[J]. E-Science Technology & Application, 2015(4): 54-59.)

[40] 乔猛,刘慧君,梁光辉. 词义层级上的专家系统问题相似度计算优化[J]. 信息工程大学学报,2018,19(4):67-72. (Qiao Meng, Liu Huijun, Liang Guanghui. Similarity calculation optimization of expert system at lexical level[J]. Journal of Information Engineering University, 2018, 19(4): 67-72.)

[41] 任晓龙,吕琳媛. 网络重要节点排序方法综述[J]. 科学通报,2014,59(13):1175-1197. (Ren Xiaolong, Lü Linyuan. Review of ranking nodes in complex networks[J]. Science China Press, 2014, 59(13): 1175-1197.)

[42] 潘锋. 特征提取与特征选择技术研究[D]. 南京:南京航空航天大学,2011. (Pan Feng. Research on feature extraction and feature selection[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2011.)

[43] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008(10): P10008.

[44] Barabási A-L, Pósfai M. Network science[M]. Cambridge: Cambridge University Press, 2016.

[45] 熊伟华,张其凡.《四库全书总目》之提要与书前提要的差异[J]. 学术研究,2006(7):134-138. (Xiong Weihua, Zhang Qifan. The differences between *Siku Quanshu Zongmu* Tiyao and Shuqian Tiyao[J]. Academic Research, 2006(7): 134-138.)

[46] 甄洪永. 明初经学研究[D]. 济南:山东大学,2009. (Zhen Hongyong. Early Ming Dynasty's study of the classics[D]. Jinan: Shandong University, 2009.)

[47] 段莹. 宋代目录学研究[D]. 郑州:郑州大学,2006. (Duan Ying. Research on Chinese classical bibliography in Song Dynasty[D]. Zhengzhou: Zhengzhou University, 2006.)

[48] Manning C P, Raghavan, H, Schütze. Introduction to information retrieval[M]. Cambridge: Cambridge University Press, 2009: 165.

[49] Griffiths T L, Steyvers M. Finding scientific topics[C]//Proceedings of the National Academy of Sciences 101(Suppl_1), 2004: 5228-5235.

[50] Juan C, Tian X, Tao L J. A density-based method for adaptive LDA model selection[J]. Journal Neurocomputing. 2009, 72(7-9): 1775-1781.

[51] Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures[C]//Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. New York: ACM, 2015: 399-408.

李　惠　上海图书馆(上海科学技术情报研究所),南京大学信息管理学院博士后。上海 200031。

陈　涛　中山大学信息管理学院副教授。广东 广州 510006。

侯君明　上海古籍出版社编辑。上海 200001。

刘　丁　天津工业大学计算机科学与技术学院讲师。天津 300061。

朱庆华　南京大学信息管理学院教授。江苏 南京 210023。

刘　炜　上海图书馆(上海科学技术情报研究所)研究员。上海 200031。

(收稿日期:2020-09-29;修回日期:2021-06-25)