面向图书采选的语义化查重策略*

漆月[1] 石璐[2]
(1. 西南大学图书馆，重庆 400715；2. 上海诺基亚贝尔股份有限公司研发部，北京 100010)

摘要：现有图书馆采选查重系统只能实现对书号、题名的重复检查，但图书出版同质化日益严重，针对异号相似图书查重困难的问题，构建基于自然语言处理技术的查重策略。首先选择主题词、内容提要和目录作为图书内容特征的指标进行建模，利用Word2Vec和WMD技术实现不同长度特征文本语义化相似度计算；然后，采用AHP方法计算特征相似度的权系数，得到图书相似度的综合评价指标；最后以西南大学图书馆数据为实验对象，验证查重策略的可行性。

关键词：采选查重；文本相似度；语义分析；评价指标体系
中图分类号：G25
DOI：10.3772/j.issn.1673-2286.2019.11.008

图书同质化是指相同或不同出版社出版的不同图书在内容上基本相同，甚至改变原书名和封面再次出版的现象[1]。当今的出版市场图书同质化现象严重[2]，导致馆藏图书的重复率逐年增加，尤其是名著类、教科类图书，多数高校图书馆都存在较严重的内容重合现象。以西南大学图书馆为例，仅C语言教材类图书的馆藏就在200种以上，且大部分借阅量为0。同质化图书的采购不仅影响图书馆的藏书质量，而且造成不必要的资源浪费。然而，逐年递增的图书出版量正在不断增加图书查重的工作量和复杂度，图书采选人员难以在海量的征订目录中深入了解每一种图书的具体内容。因此，图书馆需要寻求一种新的技术和策略，对同质化图书进行高效率的自动化判别，以减轻采选人员的工作负担，提升馆藏建设质量。

1 图书采选查重工作现状

图书查重是在拟购书单中检查是否存在已入藏的相同图书，避免因重复采购造成资源浪费和馆藏臃肿。早期的查重方法主要以ISBN号为检索点，但并不能应对ISBN编号不规范或一书多号等问题[3]。虽然改进后的图书管理系统增加了对题名、著者、出版日期、出版社等多种书目数据的排列组合与匹配功能，对查重工作起到一定的辅助作用，但基于字符串的模糊匹配方式，对于同质化图书的识别能力非常有限。

对于图书同质化问题，已有一些学者进行了研究和探讨。蒋鸿标等[4]根据外部特征将同质化图书分为显性同质化和隐性同质化，并提出针对不同特征的采访控制措施。陆文静[5]提出同质化现象是因编辑把关职能失灵造成的，并呼吁图书编辑善用职能改善这一情况。但这些措施都需要人工完成，要求相关人员利用经验和常识进行鉴别，效率较低，且容易存在遗漏、误检等情况。针对这一问题，本文提出一种基于自然语言处理技术的图书查重策略，针对图书特征进行语义相似性比较和重复度综合评价，能有效识别书目表现不同但实质内容相同或相近的图书，对现有查重方式起到补充和完善作用。

2 语义分析技术概述

语义相似度本身是一个抽象的概念，基于二进制规则运行的计算机难以对两个文本的内容相似度直接

*本研究得到重庆市教育科学“十三五”规划2019年度规划课题“面向碎片化学习的生态型智慧教学平台构建研究”（编号：2019-GX-306）资助。

进行量化计算。传统的字面匹配方式仅对字符本身进行比较，但对于同义词或者多义词的情况则无能为力。因此，需要采取某种方法将文本的语义转化为数学模型，通过计算模型之间的差异来度量两个内容的相似度，这就是语义相似性计算的基本思想。

2.1 词向量简介

词语向量化就是将词语用向量来进行表示，以便计算机识别和理解。早期的自然语言处理系统通常利用词语在词典中的位置对词语进行编码，如假设词语“儿童”“小孩”“苹果”分别出现在词典的第3位、第6位和第9位，则可以分别表示为：

儿童——[0 0 1 0 0 0 0 0 0…]

小孩——[0 0 0 0 0 1 0 0 0…]

苹果——[0 0 0 0 0 0 0 0 1…]

每个向量的长度为词典中词语的个数，这种编码方式被称为独热编码（one-hot representation）[6]，其优点是简单、直观，但容易造成维度灾难，且不能解释词语之间的关联性。

针对这一问题，Hinton[7]提出了词向量的分布式表示（distributed representation），其核心思想为：词语的语义是通过上下文信息来确定的，即相同语境出现的词，其语义也相近。具体实现方式是用统计学的方法对一个大型文本（语料库）进行训练，计算词语在给定上下文中出现的概率，并以概率值为元素在一个较低的向量空间中映射该词语。上文中的3个词语可以用四维词向量表示为：

儿童——[0.99 0.99 0.05 0.7]

小孩——[0.99 0.05 0.93 0.6]

苹果——[0.02 0.01 0.98 0.5]

因为在训练过程中考虑了词语的上下文，使得分布式表示生成的词向量带有了语义信息，如“儿童”和“小孩”词义相近，在向量空间中的距离也更近，因此只需计算出两个词向量的空间距离（通常用余弦距离表示），即可表示其语义上的量化相似度。

2.2 基于Word2Vec训练的词向量

生成词向量的分布式表示的方法有很多，其中Word2Vec是目前比较主流的词向量生成工具之一。Word2Vec技术是谷歌公司在2013年开源的自然语言处理工具，它利用人工神经网络对语料库进行训练，以实现词向量计算。Word2Vec分为CBOW和Skip-gram 2种训练模型，其中CBOW模型的训练输入是前后邻近的若干个词语，输出是最可能出现在这几个词语中间的词语，适用于小型数据库；而Skip-gram则相反，输入为某特定词语，输出为该词语前后的邻近词语，更适合大型语料库的计算[8]。设定上下文词语个数为2，2种模型如图1所示。

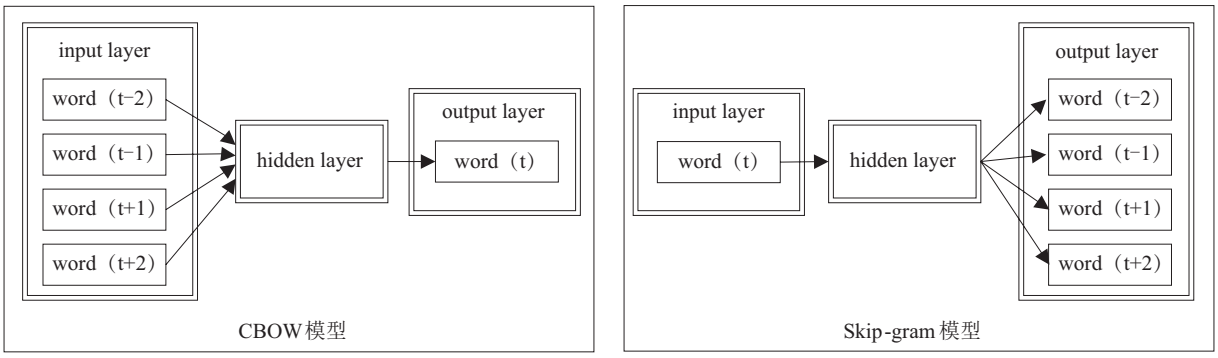

图1 Word2Vec的2种训练模型

可以看出，Word2Vec是一个三层神经网络，分别为输入层（input layer）、隐藏层（hidden layer）和输出层（output layer），模型的输入和输出都是词语的独热编码，在隐藏层中完成输入到输出的训练。训练过程中将得到一个权重矩阵，用特定词语的独热编码乘以该矩阵，即可得到该词Word2Vec训练的词向量。

3 图书的语义化查重

3.1 总体思路

虽然直接对全文内容进行比较得到的结论最为精确，但一本图书动辄数十万字，对系统的吞吐量、稳定

性要求较高，难以保证数据的处理效率和存储需求。本文认为，书目信息中的主题词、内容提要和目录3个字段足以表达图书的基本内容特征，因此选择以上3种数据作为图书特征，实现图书的语义化特征模型构建。整个同质化查重策略的总体思路是：首先根据选择的特征字段设定图书特征模型，然后对所有馆藏图书进行建模，得到馆藏特征库，在进行查重工作时，将征订书单中的图书与特征库进行语义化比较，根据相似度计算结果判定待选图书与馆藏的同质化程度。研究中以Python为实验环境，采用图书馆馆藏书目数据作为语料库构建向量词典，根据词典查询得到特征内容的向量模型，最后通过计算2种特征模型间的相似度作为图书同质化程度的量化评价。

3.2 构建向量词典

3.2.1 语料库预处理

本文选择馆藏书目数据为原始语料库以增加图书领域语义分析的准确性，通过对书目marc的结构分析，选择题目、内容提要、一般附注和主题词4个字段的数据进行抽取。由于marc中未包含目录信息，因此需要从第三方平台提供的接口获得数据，如京东联盟、当当开放平台等都提供了返回图书目录的API。实验中以西南大学图书馆书目管理系统为主要数据来源，将每种馆藏书目信息保存为一行文本，最后得到关于馆藏书目数据的语料库。对语料库的文本内容进行分词、标点符号过滤、停用词过滤、词性过滤后，即可得到待训练的语料文件。

分词处理是中文自然语言处理过程的首要步骤，因为中文以汉字为单位，词语间没有明显的分割符号，必须将一段连续的汉字分割成有意义的词语，才能用于语义分析处理[9]。本文采用Python的第三方模块jieba工具实现分词操作。jieba工具是基于Python开发的开源中文分词组件，同时还提供词性识别功能，因此在分词操作的同时可完成停用词和词性过滤操作。停用词和词性过滤的目的是删除在语义分析中作用不大的功能词，如关联词、副词、助词等。本文结合实际经验，主要保留了名词、动词和形容词相关词性的词语。另外停用词过滤采用了百度停用词表，根据词表对文本中出现的停用词进行删除。

3.2.2 生成向量词典

本文调用了Python的第三方模块gensim提供的函数gensim.models.Word2Vec对处理好的语料库进行学习，生成面向书目数据的词向量，其中训练窗口设置为5（分析上下文的前后各5个词），词向量维度设置为100，根据数据情况选择Skip-gram模型进行训练，最终输出构建好的向量词典。通过采集得到图书数据1 787 412条，得到训练好的词向量283 552个，向量词典文件输出形式如图2所示。

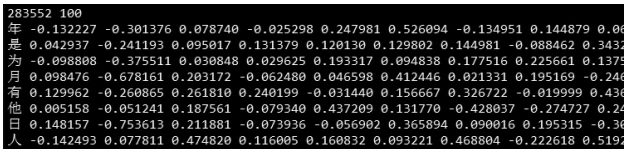

图2 向量词典文件输出形式

词典中第一行标注了词典的词汇量和词向量的维度，后面每一行记录了一个词语的向量值。将向量词典以文件形式保存后，使用时直接加载即可，不需要重复训练，同时也可以根据馆藏的增长情况在必要时进行更新。

3.3 图书特征模型设计

由于选定的图书特征字段文本长度不一，本文采用结构体的形式进行图书特征模型描述，以“keyword”（主题词）、“abstract”（内容提要）和“catalogue”（目录）为结构体的成员变量，以ISBN号为每个实例的唯一性标识。每个成员变量都是一个以词语为元素的字符串型数组，其中，“keyword”可以直接存储，其他成员则需要对文本进行分词和停用词、词性过滤操作后再将得到的词语列表保存为数组，针对“catalogue”，还过滤了章节编号、前言、参考文献、附录等对语义分析无意义的文字。对于“abstract”和“catalogue”字段的长文本处理方式是：首先进行文本分词处理后，对重复出现的词语进行去重，去重前需要记录词语出现的频率，作为词语权重用于后续工作，词频的计算方法是用该词在本文出现的频率除以该词在所有文本中出现的频率，最后将数组元素以“词语/权重”的形式存储。以《C语言程序设计》为例，其实例化后的图书特征模型形式如图3所示。

<structure>
Book: book1
int isbn=9787111617945
array keyword=['C语言', '程序设计']
array abstrcat=['程序/0.086956', '系统/0.086956', '结构/0.086956', '语言/0.043478', '特性/0.043478', '方法/0.04,...]
array catalogue=['运算符/0.130952', '函数/0.119047', '指针/0.089285', '语句/0.089285', '类型/0.077380', '表达式...]

图3 图书特征模型实例

3.4 特征相似度计算

完成图书特征建模后，即可利用Word2Vec构建的向量词典实现语义化相似度计算。在语义分析过程中，长文本与短文本的处理方式不同，需要分别进行处理。这里首先讨论计算2种图书的主题词相似度方法，具体步骤为：①遍历“keyword”的所有元素，在向量词典中查找每个元素的词向量；②计算词向量在每个维度上的平均值，作为“keyword”的特征向量；③计算2个特征向量的余弦距离作为2个特征的相似度评价值，余弦距离记为cosDict（***K***，***K***’）。其中，***K***与***K***’分别表示两种图书的“keyword”特征向量。

由于“keyword”仅有词语组成，词语间基本不存在关联性，所以只需取每个数组元素的词向量平均值作为图书主题词的特征向量，即可完成相关计算。在短文本的语义分析中，这种做法不会产生太大误差，但是在长文本的相似度计算时，忽略上下文的语义关联则会影响结果的精确度。因此，在面向“abstract”“catalogue”的长文本计算时，本文根据词语在上下文中的语义关系为每组距离进行加权，再对距离进行求和得到两个文本向量的空间距离，以得到更加准确的长文本语义相似度结果，记为txtDict（T，T’）。其中，T与T’分别表示两个长文本类型的特征数组，而距离权重通过WMD算计得到。WMD是一种基于Word2Vec技术的文本距离算法，它通过乘以计算文本T中的词语转移到文本T’中所需要的最小代价来衡量两个文本的相似度[10]。

3.5 图书同质化查重策略

整个同质化查重策略的流程如图4所示，首先以图书馆的书目数据为语料库进行Word2Vec训练得到向量词典，同时通过对馆藏图书的特征建模生成图书特征库。在图书采选时，将征订书单中的图书与特征库的所有图书分别进行特征相似度计算，从而作出同质化程度判定。

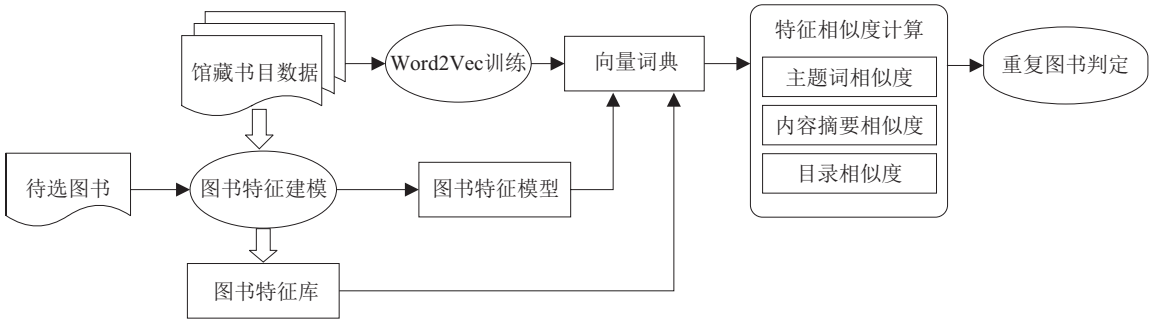

图4 图书同质化查重流程

同质化判定将主题词、摘要和目录3种特征相似度作为评价指标，通过权重设置综合3种评价指标得到图书的相似度判定如公式（1）。

$$sim(B,B)=\alpha_0\times cosDist(B.keyword,B'.keyword)+\alpha_1\times txtDist(B.abstract,B'.abstract)+\alpha_2\times txtDist(B.catalogue,B'.catalogue) \quad (1)$$

其中B和B’是2种图书的特征模型，α_0、α_1、α_2为3个特征相似度的权系数。

公式中的权系数的计算采用AHP来实现[11]，具体流程如下。

步骤一：经多专家在[1~9]数值区间内对评价指标的重要程度赋值，其中“1”表示2种指标相比具有相同的重要性，“9”表示前者比后者极端重要。将主题相似度、摘要相似度、目录相似度分别记为C1、C2、C3，得到3种指标两两比较的判断矩阵***P***，如表1所示。

表1 评价指标两两比较判断矩阵

	C1	C2	C3
C1	1	1/3	1/6
C2	3	1	1/4
C3	6	4	1

其中，Ci∶Cj≥a_{ij}，a_{ij}表示Ci与Cj相比对相似性判定的重要程度，$a_{ij}=1/a_{ji}$。

步骤二：计算表1中判断矩阵的特征向量$\boldsymbol{W}$，其中$w_i=\sqrt[3]{(a_{i1}\times a_{i2}\times a_{i3})}$，$\boldsymbol{W}=(w_1, w_2, w_3)=(0.41, 0.91, 2.71)$，进行归一化处理后，$\boldsymbol{W}=(0.10, 0.23, 0.67)$，同时，可以计算出最大特征值λ=3.09。

步骤三：进行一致性检验分析，检验判断矩阵的逻辑正确性。检验指标用一致性比率CR值来分析，CR值越接近0，表明判断越正确。计算公式为 $CR=\frac{CI}{RI}$，其中CI是一致性指标，$CI=\frac{(\lambda-3)}{(3-1)}$；RI表示随机一致性指标，通过查找随机一致性参照表得到RI=0.58。计算得到CR=0.07，证明判断矩阵具有满意的一致性。最后得到的图书相似度综合评价指标如公式（2）。

$$sim(B, B')=0.10\times cosDist(B.keyword, B'.keyword)+0.23\times txtDist(B.abstract, B'.abstract)+0.67\times txtDist(B.catalogue, B'.catalogue) \quad (2)$$

4 实验验证

为验证查重策略的有效性，本文利用西南大学图书馆20180129批次的征订书单进行查重检验。该批次已经经过采访人员处理，对相似性图书进行了标注，便于统计识别结果。实验中检测阈值设置为0.7，即将相似度大于70%的图书判断为重复馆藏。

实验操作在完成ISBN查重之后进行，以该馆现有的题名、责任者等书目查重方式为实验参照，分别采用准确率、召回率和*F*值3种指标对实验结果进行评价。统计结果如表2所示。

表2　查重结果对比

查重方法	准确率	召回率	*F*值
书目查重	47.12%	40.54%	43.58%
语义化查重	82.33%	74.55%	78.25%

可以看出，基于基础书目数据的查重方式对同质化图书识别率较低，而通过语义化查重能够使识别成功率得到明显提升。

5 结语

图书市场需求量的增长带动了图书出版行业的快速发展，同时也导致图书出版中的同质化现象日趋严重。现有的图书管理系统的采选查重功能并不完善，难以对图书实质内容的重复性进行判断，而人工查重不仅效率低，且容易出现漏采、错采的情况。因此，本文设计了一种针对图书内容的语义化查重策略，通过自然语言处理技术对图书的特征信息进行相似度计算，帮助采编人员提升图书查重效率和准确率，从而有效减轻馆藏重复现象，优化图书馆馆藏结构。

虽然在研究过程中仅选择了主题词、内容提要和目录3种数据作为图书内容相似度评价指标，但在延伸研究中，也可以采用本文的思想增加更多的特征信息进行计算，如增加对作者、出版社等因子的分析，能够对图书再版、连续出版等情况进行更详细地区分。

参考文献

[1] 张圣阳. 图书同质化的思考 [J]. 黑龙江史志，2014（15）：166-167.

[2] 张岩. 传统文化图书出版中的问题与编辑责任刍议 [J]. 出版发行研究，2018（7）：62-64.

[3] 肖婷. 从ISBN的唯一性谈中文图书采访的查重 [J]. 图书馆工作与研究，2013（5）：85-87.

[4] 蒋鸿标，吴为民. 中文图书出版同质化采访控制研究 [J]. 上海高校图书情报工作研究，2018，28（2）：59-63.

[5] 陆文静. 善用编辑把关职能减少图书同质化现象 [J]. 出版参考，2016（9）：29-30.

[6] 郑啸，王义真，袁志祥，等. 基于卷积记忆神经网络的微博短文本情感分析 [J]. 电子测量与仪器学报，2018，32（3）：195-200.

[7] HINTON G E. Learning distributed representations of concepts [EB/OL]. [2019-09-10]. https://pdfs.semanticscholar.org/cef4/692d7d5f33b0d819079642c69778497415e6.pdf.

[8] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [J]. Computer Science, 2013（1）：28-36.

[9] 许峰，张雪芬，忻展红. 基于深度神经网络模型的中文分词方案 [J/OL]. 哈尔滨工程大学学报：1-6 [2019-06-30]. http://kns.cnki.net/kcms/detail/23.1390.U.20190527.1323.004.html.

[10] 官赛萍，靳小龙，徐学可，等. 基于WMD距离与近邻传播的新闻评论聚类 [J]. 中文信息学报，2017，31（5）：203-214.

[11] 王凌峰，姚依楠. 主观线性加权评价问题的新方法：中位数层次分析法 [J]. 系统科学学报，2018，26（1）：96-99.

作者简介

漆月，女，1984年生，硕士，馆员，研究方向：图书情报，E-mail：qqt.123@163.com。
石璐，女，1984年生，工程师，研究方向：人工智能、大数据技术。

Semantic Duplicate Checking Strategy for Book Acquisition

QI Yue[1] SHI Lu[2]
(1. Southwest University, Chongqing 400715, China; 2. Nokia Shanghai Bell Co. Engeneering Department, Beijing 100010, China)

Abstract: The existing system of library acquisition and duplicate checking can only work with same ISBN number or title. But in the case of serious homogeneity of book publishing, it is difficult to filter out books with similar contents, a method of book semantic duplication checking based on natural language processing technology is presented to solve this. Firstly, subject words, abstracts and catalogues are chosen as the evaluation elements to build model with library. Then, calculate the semantic similarity of context with Word2Vec and WMD, get the weight of similarity by AHP method. Then get comprehensive evaluation of book similarity. Finally, verify the duplication checking strategy with the library data of Southwest University.

Keywords: Book Duplicate-Checking; Context Similarity; Semantic Analysis; Evaluation Index System

（收稿日期：2019-10-13）

书 讯

《汉语主题词表》

《汉语主题词表》自1980年问世以后，经1991年进行自然科学版修订，在我国图书情报界发挥了应有作用，曾经获得国家科学技术进步二等奖。为适应网络环境下知识组织与数据处理的需要，由中国科学技术信息研究所主持，并联合全国图书情报界相关机构，自2009年开始进行重新编制工作，拟分为工程技术卷、自然科学卷、生命科学卷、社会科学卷四大部分逐步完成。目前工程技术卷和自然科学卷已出版。

《汉语主题词表（工程技术卷）》共收录优选词19.6万条，非优选词16.4万条，等同率0.84，在体系结构、词汇术语、词间关系等方面进行了改进创新。《汉语主题词表（自然科学卷）》共收录专业术语12.4万条，包含数学、物理学、化学、天文学、测绘学、地球物理学、大气科学、地质学、海洋学、自然地理学等学科领域，收词系统、完整，语义关系丰富、严谨，每条词汇都有相应的学科分类号表现其专业属性，并与同义英文术语对应。同时，建立《汉语主题词表》网络服务系统，提供术语查询、文本主题分析、知识树辅助构建等服务。《汉语主题词表》可用于汉语文本分词、主题标引、语义关联、学科分类、知识导航和数据挖掘，是文本信息处理及检索系统开发人员不可或缺的工具。

《汉语主题词表（工程技术卷）》已于2014年由科学技术文献出版社出版，分为13个分册，总定价3 880元。《汉语主题词表（自然科学卷）》已于2018年5月由科学技术文献出版社出版，分为5个分册，总定价1 247元。两卷均可分册购买。