科技项目申报书查重方法研究

王 东 王 飘 江俊鹏 李 青 徐晨阳
（中国科学技术信息研究所，北京 100038）

摘要：开展面向科技项目申报书的查重方法研究，对于推进学术诚信建设、营造风清气正的科研环境具有重要意义。目前，关于科技项目申报书的查重研究仍处于起步阶段，针对存在的查重系统架构不明确、查重算法准确率较低等问题，构建一套涵盖科技项目申报书数据处理、分布式任务、查重算法模块与查重报告生成的系统模型，并在查重算法方面提出基于DSSM架构的相似度检测算法模型。实验结果表明，该查重系统能够实现较高的查重准确率和查重效率，能够在科技项目申报书查重方面发挥积极的作用。

关键词：科技项目申报书；DSSM架构；文本相似度；查重算法；查重系统
DOI：10.3772/j.issn.1674−1544.2022.05.004 **CSTR：**15994.14.issn.1674.1544.2022.05.004
中图分类号：TP391 **文献标识码：**A

Research on the Duplicate Checking Method for Scientific and Technical Project Applications

WANG Dong, WANG Piao, JIANG Junpeng, LI Qing, XU Chenyang
(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: It is of great significance for promoting the construction of academic integrity and creating a clean and positive scientific research environment to carry out the research on duplication checking methods for the declaration of scientific and technical projects. At present, the research on duplicate checking of scientific and technical project application is still in its infancy, and there are problems such as unclear duplicate checking system architecture and low accuracy of duplicate checking algorithm. To solve these problems, this paper designs and implements a system model covering data processing of scientific and technical project declaration, distributed tasks, duplicate checking algorithm module and duplicate checking report generation, and proposes a similarity detection algorithm model based on DSSM architecture in duplicate checking algorithm. The experimental results show that our duplicate checking system can achieve high duplicate checking accuracy and efficiency, and we believe that it can play a positive role in duplicate checking of scientific and technical project declaration.

Keywords: Declaration of Scientific and Technical Projects, DSSM architecture, text similarity, duplicate algorithm, duplicate checking

作者简介：王东（1987—），男，中国科学技术信息研究所助理研究员，研究方向为数据挖掘、知识图谱、数据可视化；**王飘**（1992—），女，中国科学技术信息研究所助理研究员，研究方向为计算机应用、信息系统建设；**江俊鹏**（1992—），男，中国科学技术信息研究所助理研究员，研究方向为信息资源管理；**李青**（1991—），男，中国科学技术信息研究所工程师，研究方向为云存储、数据集成；**徐晨阳**（1992—），男，中国科学技术信息研究所助理研究员，研究方向为信息安全、信息系统运维（通信作者）。

收稿时间：2022 年 7 月 8 日。

党的十九大以来，党中央坚持把科技创新摆在国家发展全局的核心位置，科研工作投入稳步增长。然而，随着各类科技项目的数量不断提升，科技项目重复立项的问题日益突出，针对科技项目进行查重已成为科技项目管理的一个重要的技术环节。而科技项目申报书是科研团队为获得各级科技管理部门对拟申报项目的研究经费许可、按照标准格式填写的项目申报文档，对拟申报项目从进度安排、研究内容、预期效益、组织实施等部分进行综合论述，因此填写并提交科技项目申报书是科技项目申报和立项过程中必不可少的一环[1]。查重科技项目申报书，对于避免科技项目重复立项有着至关重要的作用。本文针对科技项目申报书提出了一种基于DSSM架构的相似度检测算法模型，并在此基础上构建了一套查重系统，希望能够在科技项目查重过程中发挥积极的作用。

1　研究现状

通过文献调研发现，在国外科技项目的申报和评审主要通过同行评议的方式进行，尚未发现明确的项目查重要求。而在国内，面向科技项目以及其他各类文档的查重方法则有较多研究。刘如等[2]结合国内外文本查重技术的进展情况，设计了一种面向科研项目申报材料查重的解决方案，深入探索了将深度学习与文本查重相结合的方法。黄思颖等[3]针对现有查重系统存在的性能较低、准确率不高以及扩展性不好等问题，提出了一套基于SolrCloud的分布式科技项目查重系统，在实时性能和扩展性等方面有着较好的表现。陶秀杰等[4]针对企业环境下科技项目的查重需求，提出了一种结合新词发现和中国知网KBase工具的查重方法。高爽等[5]针对环境影响评价报告书中可能存在的抄袭行为，创建了一套基于simhash的智能查重方法，用于辅助报告书的初步筛查和技术复核。刘玉林等[6]针对项目重复招标的问题，提出了一套基于潜在语义索引的文本相似度检测平台，可以快速、准确地分析待检测文档与海量在库文档的相关性，在一定程度上规范项目招标采购管理工作。但是，鲜有对科技项目申报立项时进行查重的专门研究。本文将重点对科技项目申报书的查重进行研究，构建查重系统，以避免科技项目重复立项，减少科研经费的损失，加强科研诚信建设。

2　查重系统架构

2.1　总体架构

本文提出一个4层查重系统总体架构，分别是查重应用层、数据处理层、基础组件层、运维管理层，如图1所示。

（1）运维管理层。运维管理层是总体架构的底层，主要功能是管理、调度、协调运维资源，控制维护各类容器，保障多用户空间资源的分配与隔离。

（2）基础组件层。基础组件层汇聚适用于计算、存储、查询、消息队列的多种大数据组件，可有效使用运维管理层的运维资源，向数据处理层提供基础服务。

（3）数据处理层。数据处理层是总体架构的数据中心，调用基础组件层内部的各类大数据组件，开展科技项目申报书相关数据的采集、转换、加载、分析计算，向上对查重应用层做好数据支撑。

（4）查重应用层。基于数据处理层提供的数据，通过运维管理层统一的任务调度体系开展应用容器化管理控制，提供查重任务管理、文本查重、查重报告生成等查重应用服务。

2.2　工作流程

该查重系统的工作流程分为数据预处理、查重算法执行、查重报告生成3个主要步骤，如图2所示。

2.2.1　数据预处理

数据预处理模块是一个前置任务层，主要用来处理巨大的申报书数据信息。为满足上层模块需要的数据定位、数据处理、数据存储等业务需求，该查重系统提取了数据方面的基础工作作为一个模块封装，为上层服务提供工具。数据预处理模块所有功能都基于高效地使用HDFS

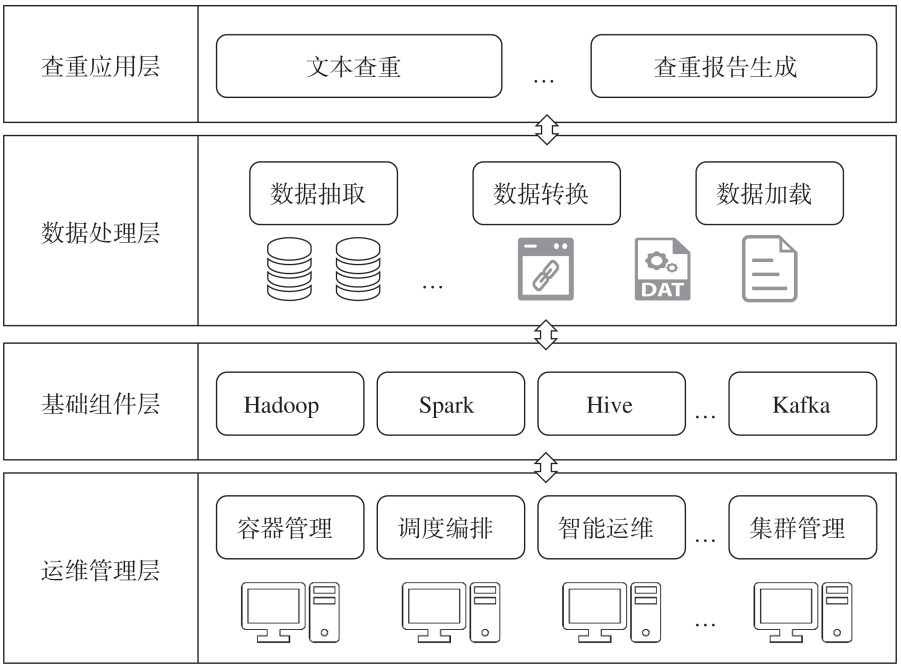


图 1　查重系统总体架构

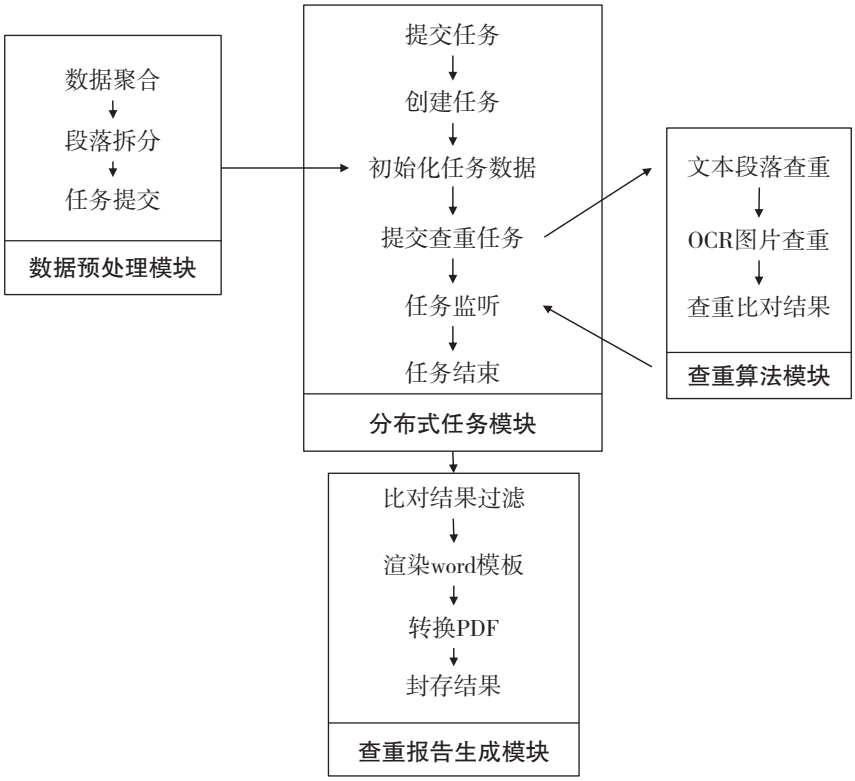


图 2　查重系统工作流程

（Hadoop分布式文件系统）提供的数据存储能力。数据预处理模块主要实现以下 4 种功能。

（1）基础数据库建立。通过科技管理信息系统提供的接口，将可作为查重依据的申报书PDF统一标识后存入HDFS数据系统。在此统一标识基础上生成结构化数据对象。对象记录了PDF类型、状态、编号、版本号、提交时间等信息，便于后续查重算法调用。该功能具有容错和恢复机

制，保证大数据情况下的数据完整性，具有可靠性高、可用性强等特点。

（2）新数据入库。因为查重依据数据会动态增加、修改与删除。该模块可以修改HDFS数据系统下的PDF申报书数据。对于科技部新入库申报书数据，通过接口下载后，该模块会与本地HDFS数据系统中已保存的PDF申报书进行比对，确定PDF是否已在本地库、新申报书PDF是否修改等，以此判断新申报书是否应该入库。如果新申报书覆盖了原有申报书，表示修改该申报书版本号并更新提交时间。以上服务保证了该系统的查重数据库是与科技部数据一致的，同时不会出现数据重复。

（3）PDF申报书文本化。PDF存储可以使用原始数据格式，但是在后续的查重算法模块、报告生成模块却无法直接使用PDF申报书原始数据。因此，在其他模块调用时，该项服务会动态切入，将从HDFS文件系统中调用的PDF申报书转化为文本数据，提供给其他模块使用。

（4）特殊格式申报书处理。大部分申报书原始格式是PDF，但也有以Excel或表格提交的申报书。对于这些申报书，系统将统一在数据预处理层进行规范化处理，以避免因特殊格式而导致在后续其他模块中需要再添加重复逻辑代码的处理。

2.2.2　查重算法执行

查重算法模块是查重系统的核心技术模块。该模块将需要查重的申报书视为多个段落，通过分布式任务模块，将每一个段落与查重数据库中的其他申报书进行查重。在查重过程中，将重复率大于指定阈值的段落信息记录，结合重复率与其他信息封装成结构化数据，保存在对象中，在所有查重任务完成后，将查重数据提供给生成申报书模块使用。

第一步：查重总体结论。

第二步：确定每章重复率。查重系统是以段落为单位，分布式进行查重，因此每个段落会与查重数据库中的所有段落进行算法比对，比对结束后，会得到N个该段落在查重算法检测后认定存在高重复度的范围区间α。这些区间经过相并操作之后得到的区间就是该段落的重复区间，除以段落总长度*len*就可以得到该段落的相似比*k*，如式（1）所示。

$$k = \frac{U_i^N a_i}{len} \tag{1}$$

第三步：提供相似样本列表。在查重算法过程中，以段落为查重算法基础单位，但是为了便于人工分析相似申报书样本，该模块同时会以申报书为单位，记录查重申报书与其他申报书的总体相似度。在查重算法结束后，将与查重申报书相似度最高的若干申报书记录，作为相似样本提供查询与分析。

第四步：提供主体数据列表。从检测目标申报书中与对比样本内容中选择具体的重复内容记录，给报告生成模块提供生成查重报告。

2.2.3　查重报告生成

在对申报书进行的查重任务完成之后，便可以使用查重过程中关键性信息生成的结构化数据来构成最后的查重报告。报告生成模块使用了渲染+动态生成的方法来生成Word文档，之后将Word文档转化为PDF文档报告。对于该报告，将封装留存并返回给使用者进行查询与分析。具体报告生成流程如图3所示。

3　查重算法研究

3.1　算法架构

查重算法文本模型结构上是先通过最大字符串匹配的方式找到高相似度的段落，然后根据深度学习算法模型来判断两个段落的语义相似度作为最终依据[7]。这样的设计结合了最大字符串匹配的快速查找与深度学习模型判断重复的高准确率。本文在已有项目申请书语料之上，实现与改进语义相似度算法，最终经过实验对比，选择了DSSM（Deep Structured Semantic Model）架构作为主体框架，使用Transformer替换DSSM表现层的算法模型。算法模型架构如图4所示，其中DSSM原模型表示层使用DNN来提取语义特征。本文主要研究在表示层上结合最新语义提取表示模型如Transformer模型，以得到较好的算法效

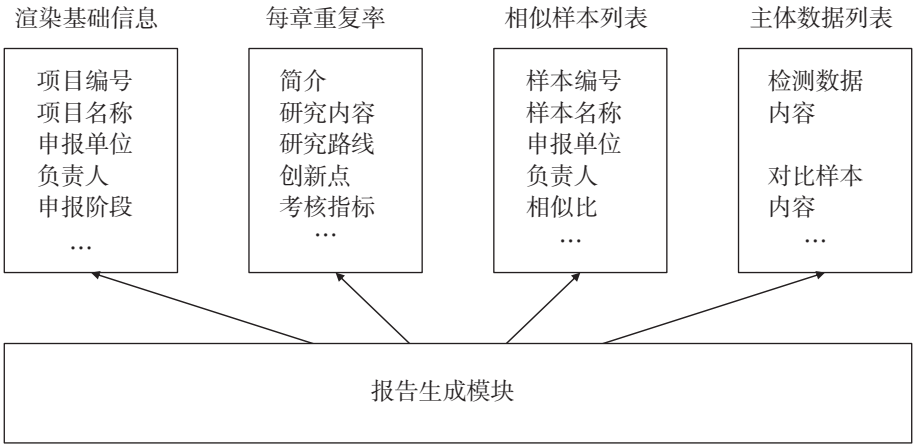


图 3　报告生成模块流程

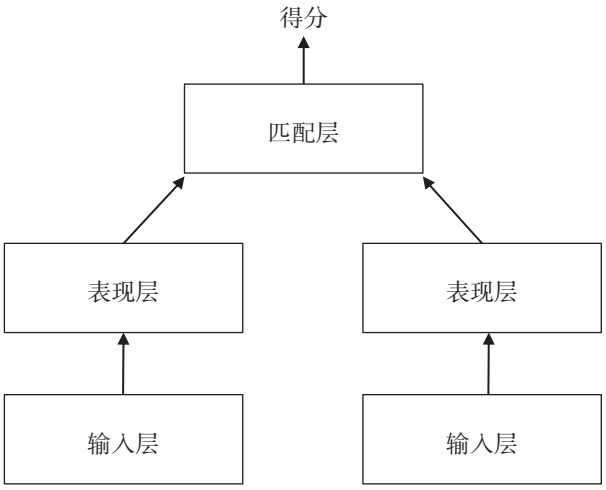


图 4　查重算法模型架构

果。同时，本文将研究注意力机制在该架构上的使用。通过在匹配层加入全局注意力机制，使模型能够关注到语句中最重要的语义信息。

3.2　模型输入层

首先，使用分词模型如jieba、pkuseg等对查重系统数据预处理模块提供的科技项目申报书段落句子，进行分词和词过滤处理，构建初始科研文本A与科研文本B。然后，使用大数据量训练好的Word2Vec模型将文本向量化，得到句子的向量矩阵，从而在学习时保留词语的语义信息。具体操作如图 5 所示。

在经过Word2Vec模型向量化之后，得到输入句子中按每一个单词先后顺序对应的有序向量列表。为了便于表示层Transformer训练方便，本文需要保证得到的两个向量列表长度一致。因为输入文本得到的词组数量有差异，词向量列表可能会出现列表长度对不齐的问题。本文采用补齐与截取的方法来处理该问题。首先设置一个词向量列表最大长度，当输入词向量列表长度小于该值时，在该词向量列表的尾部填充向量直至列表长度等于最大长度。填充向量类型有随机初始化向量或零向量，本文使用零向量进行填充。当输入词向量列表长度大于最大长度时，将超过最大长度的部分截取舍弃。补齐与截取操作的表达式如式（2）所示。

$$s = w_1, w_2, ..., w_l = \begin{cases} \left[w_1, w_2, \ldots, w_n, \vec{0}_{n+1}, \ldots, \vec{0}_l \right], \text{if } n < l \\ \left[w_1, w_2, \ldots, w_l \right], \text{if } n \geq l \end{cases} \quad (2)$$

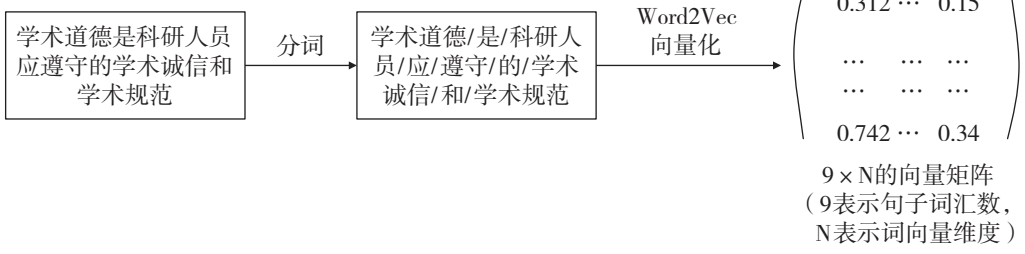


图 5　Word2Vec生成向量矩阵

在式（2）中，s表示输出的词向量序列，w_i是输出序列的词向量，n表示句子词向量个数，l表示设置的最大长度，$\vec{0}_j$表示补齐的零向量。

本文在对数据集数据进行统计分析后，将最大长度设置为32。通过以上输入层的操作，将原始文本转换为长度一致的词向量列表，然后将其输入到表示层进行训练。

3.3　模型表示层

该模型的表示层采用了基于Transformer机制的编码层来进行文本特征提取，而非原始DSSM模型所用的DNN，这是由于Transformer的注意力机制与位置编码可以比较好地保留了文本的顺序信息，同时Transformer本身拥有非常强的特征提取能力。

Transformer编码层自注意机制网络层结构由多头自注意力机制与多头注意力机制（Multi-Head Attention）两部分组成[8]。多头自注意力机制（Multi-Head Self-Attention）self-attention结构在进行运算时需要使用矩阵Q（Query）、K（Key）、V（Value），这3个矩阵是经过线性变化self-attention输入计算产生的。Self-attention在进行计算时使用了缩放点积，因此self-attention的输出表达式见式（3），其中dk是Q、K两个矩阵的列数，即向量维度：

$$Attention(Q,K,V)=softmax\frac{QK^T}{\sqrt{d_k}}V \qquad (3)$$

多头注意力机制组成如图6所示，主要思想是使用h个不同的线性变换对Q、K、V进行运算

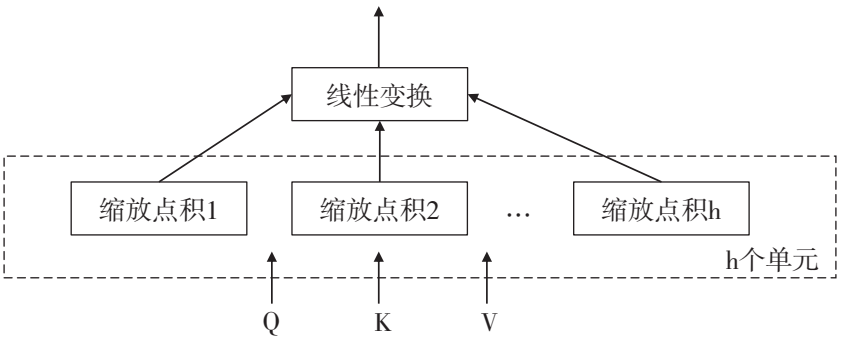


图6　多头注意力结构示意

投影，然后拼接计算结果，进行线性转换得到输出结果。这就是多头注意力的含义[9]。其计算方式如下：

$$head_i=Attention\left(QW_i^Q,KW_i^K,VW_i^V\right) \qquad (4)$$

$$MultiHead\left(Q,K,V\right)=Concat\left(head_1,head_2,\ldots,head_h\right)W^O \qquad (5)$$

通过多头注意力后就是全连接前馈网络，每个部分都添加了残差连接和归一化，本文不再展开介绍。Transformer得到的是单个词的向量表示，而查重是以段落为单位，因此本文使用全局注意力（Global-Attention）来计算两个段落的最终特征向量，如图7所示。

全局注意力模块在一个对齐向量上，对所有时间序列的隐藏层进行压缩。c_t表示上下文向量，是编码器中整体时间步的隐藏状态加权和，时间步上隐藏状态向量的维度即编码器隐藏层的神经

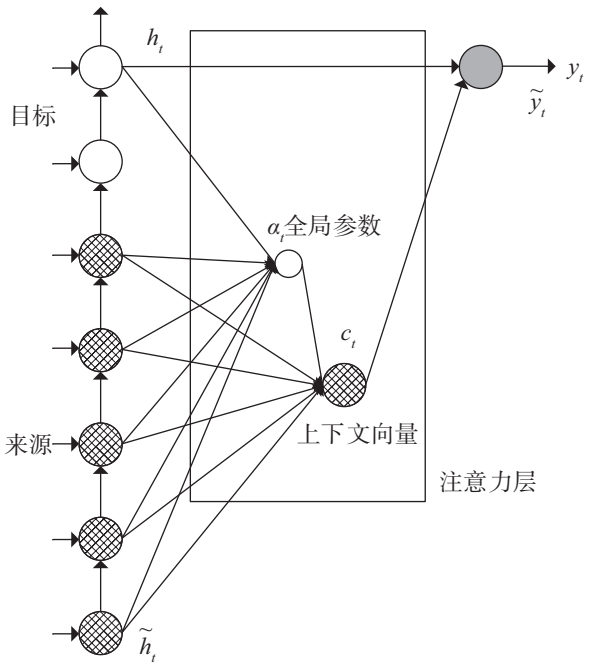


图7　全局注意力结构

元数量。其中，编码器的隐藏状态和上下文向量的维度一致。注意力权重的计算方式见式（6）：

$$\alpha_t(s)=align\left(h_t,\bar{h}_s\right)=\frac{\exp\left[score\left(h_t,\bar{h}_s\right)\right]}{\sum_{s'}\exp\left[score\left(h_t,h_{s'}\right)\right]} \quad (6)$$

在式（6）中，$\alpha_t(s)$是当前时间步状态在其他时间步状态中的占比，c_t是$\alpha_t(s)$的加权和，*score*是根据所有时间步与当前时间步的状态平均值生产相似度的预定义函数。

3.3　模型匹配层

经过模型表示层获取两个段落的语义特征后，接下来需要将两个段落语义特征进行配对，即存在相似的两组段落视为匹配。在设置的数据集对中，每个段落都只有一个相似段落作为正样本，而剩余段落都作为负样本。在匹配过程中，可以使用余弦相似度等方式计算两个段落的语义相似度。

在科技项目申报书查重研究中，可以直接将两个科研文本相似段落的语义信息向量计算其余弦值[10]。但是通过表示层获取的语义特征往往包含很多和文本主题无关的信息，而科研文本本身具有很强的指向性，每个句子信息指向明确，因此为了过滤不关键的附加信息，本文采用了注意力机制为关键信息赋以更高的权重，从而对它们施加更多的关注。实际上，本文使用了一种AOA Reader（Attention−over−Attention）模型用于匹配层，通过注意力机制的两次叠加应用计算注意力权重。如两个待匹配的科研文本记为A、B，首先以科研文本A记为Key值，由科研文本B查询科研文本A，计算出从科研文本B到科研文本A的注意力权重；再计算出A到B的权重值；然后将权重值乘以A的语义表示，再将权重值乘以B的语义表示，既可计算出注意力机制中的科研文本A表示与科研文本B表示；最后计算两者的余弦相似度[11−12]。

如本文通过DSSM表示层得到科研文本*A*的语义表示向量**Z**，科研文本*B*的语义表示向量***X***，两个向量点乘后获得关于科研文本*A*和科研文本*B*的交互信息矩阵，然后对交互信息矩阵以行和列为单位进行Softmax归一化，从而计算出科研文本*B*到科研文本*A*的注意力权重α和科研文本*A*到科研文本*B*的注意力权重*β*。将*α*与**Z**相乘、*β*与**X**相乘，即可生成基于注意力机制的表示，计算两者的余弦相似度*R*，就可以进行梯度更新来训练模型[13−14]。

$$y_Z=\alpha\times\mathbf{Z} \quad (7)$$

$$y_X=\beta\times\mathbf{X} \quad (8)$$

$$R\left(y_Z,y_X\right)=cos\left(y_Z,y_X\right)=\frac{y_Z^{\ T}y_X}{\left\|y_Z\right\|\left\|y_X\right\|} \quad (9)$$

在模型训练过程中，对于输入的两个科研文本段落集U与V，得到如下概率化定义：

$$P(\mathrm{V}|\mathrm{U})=\frac{\exp\left[\gamma R(U,V)\right]}{\sum_{B'\in B}\exp\left[\gamma R(U,V')\right]} \quad (10)$$

γ为变量经验系数，可以使归一化后的概率更平滑，U与是所有的科研文本数据集合，集合内容包括：①训练样本内部与U存在关系的全部v（称为正样本），V^+记为与U存在关联关系的v。②训练样本中内部任意一个U，在数据库里选出随机负样本（与U不相关的v），这些负样本记为V^-，集合V包含U对应的所有相关集合V^+与无关集合V^-。

$$V^-\in V$$

$$V^+\in V$$

模型使用交叉熵作为损失函数，使用Adam优化器来迭代更新模型参数。损失函数如式（11）所示：

$$L=-\log\prod_{(U,V^+)}p\left(V^+|U\right)=-\prod_{(U,V^+)}\log p\left(V^+|U\right) \quad (11)$$

4　实验与分析

对于上文提出的基于DSSM架构的查重算法，本文使用python语言实现了单独的算法模型，使用中文文本相似度计指标来运用在模型的测试中，通过实验来进行效果评估。

4.1　实验数据集

基于上文的算法模型结构，本文训练模型需要对以下两块数据集进行准备。

（1）训练词向量所需语料库。本文训练词向

量所使用的语料库，一部分来源于科研论文的摘要和结语，另一部分来源于百度百科、维基百科等平台的公开数据。在获取上述语料后，使用基于Skip−Gram方式的Word2Vec模型进行训练。

（2）DSSM模型数据集。DSSM数据集由现有公开数据集和人工构建数据集组成。其中，公开数据集为20 000条数据量的相似文本，可用于常规的语义相似度训练。人工构建数据集的科研文本来源科技部提供的科研申报书，从各个学科分类中选取了10 000条数据量的相似文本，并且经过去冗余、删除无关文本等操作，作为原始训练数据。DSSM模型数据集的划分方法如表1所示。

表1　DSSM模型数据集分布表

数据集	数据量/条	格式
训练集	25 000	Json
验证集	2 000	Json
测试集	3 000	Json

4.2　实验结果

为了验证本文查重模型的效果，在实验中与一组成熟的语义相似度计算方法开展对比实验，主要包括CBOW、DSSM、ARC−I[11]模型。在相同的输入数据基础上进行计算。这些模型对比实验结果如表2所示。

表2　模型对比实验结果

模型	准确率/%	*F*1值/%
CBOW	73.16	73.28
DSSM	74.47	74.35
ARC−I	77.52	76.46
本文模型	78.72	78.43

从表2的实验结果可以看出，本文的查重算法模型与其他语义相似度计算方法相比，准确率、*F*1值都有所提升。在这些模型中，由于CBOW针对低层次文本特征进行计算，查重效果较差。DSSM的计算效果受其表现层DNN特征提取能力影响，实验效果一般。相比CBOW、DSSM模型，ARC−I表现较好。最后，通过实验对比，证明了本文提出模型的准确率、*F*1值均高于其他语义相似度计算方法，在科技项目申报书相似度分析上具有更好的效果。

4.3　实验分析

通过构建科研文本数据集，设置实验效果较好的模型参数。本文的算法模型在与其他模型进行对比实验后，取得了最好的实验准确率与*F*1值。可以看出，使用深度学习的语义相似度计算模型，比原始语义相似度计算方法会取得更好的效果。本文提出的基于DSSM架构，使用Transformer编码层替换DSSM表现层的算法模型具有良好的效果。其原因在于，Transformer编码层考虑到了科研文本的位置信息，并且具有更深层次的特征提取能力，同时模型在匹配层使用了AOA Reader模型过滤器，充分提取交互能力，提升了计算科研项目文本相似度的准确率。

5　查重系统效果展示与分析

5.1　效果展示

本节以科技项目申报书的查重报告为例，展示上述查重系统的成果。一份完整的查重报告包括4个部分，分别为总体结论、每章内容重复率、相似样本列表、主体数据列表。

图8展示了查重报告的“总体结论”部分。在这部分主要展示某科技项目申报书的总相似比、与已立项项目相似比、主体/非主体部分相似比、与自己/他人项目相似比、与同单位/其他单位项目相似比等数据，并以热力图形式直观展示了申报书各个部分的相似程度。

图9展示了“每章内容重复率”部分。在这部分展示了申报书各章节内容的相似比。

图10展示了“相似样本列表”部分。在这部分展示了与所查项目书最为相似的若干个项目信息，并按照相似比从高到低的顺序进行排列。

从图8至图10可以看出，本文所设计的查重系统可以较好地识别出科技项目申报书的重复或相似内容，能够满足基本查重需求，具有准确、全面、可靠的特点。此外，本文提出的查重系统还具有以下优点。

（1）系统效率高，可动态扩容。由于采用了分布式架构，整个查重系统可将查重任务分配到

各个任务节点上，从而实现并行处理，相较于单机架构在效率方面有很大的提升。同时，该系统还支持动态扩容，因此在科技项目集中申报期间可通过增设节点来避免拥挤。

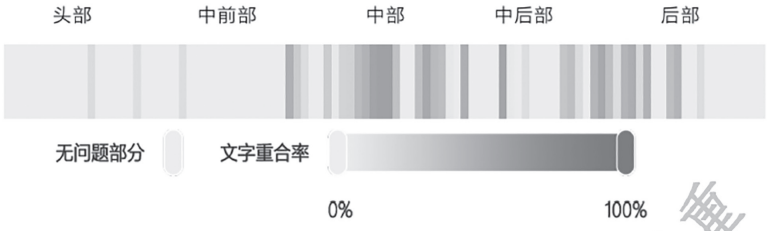
（2）系统功能灵活，可定制程度高。该系统除了最基本的全文查重功能外，还具有较多的可定制化功能，如该系统可根据申报人、单位、申报年份等属性，动态调配查重项目库，从而实现个性化的查重需求。

（3）系统查重能力强，可防止人为降重。该系统对于文本的查重策略是建立在语义相似度的基础上的，因此不仅可以识别出完全重复的句子，而且可以识别出那些虽然单词顺序不同或更换部分词语但是表达意思基本一致的相似句。

5.2　挑战及展望

本文所提出的科技项目申报书查重系统主要面对如下3个方面的挑战[15-17]。

（1）原始申报书存在格式错乱等问题。在查重过程中可以发现，许多申报人所提交的申报书存在各种各样的格式错误，如标题序号错乱或遗



总相似比： 22.43%	**与已立项项目相似比：** 19.16%
主体部分相似比： 2.82%	**非主体部分相似比：** 23.47%
与自己项目相似比： 0.00%	**与他人项目相似比：** 22.43%
与同单位项目相似比： 0.00%	**与其他单位项目相似比：** 22.43%
相似字数： 3169	**总字数：** 16032

图8　总体结论

章节	相似比
简介［主体部分］	0.00%
一、目标及考核指标、考核方式/方法［此章节不查重］	
二、课题研究内容、研究方法及技术路线［主体部分］	2.78%
（一）课题的主要研究内容［主体部分］	2.76%
（二）课题采取的研究方法［主体部分］	2.82%
三、主要创新点［主体部分］	6.18%
四、预期经济社会效益	0.00%
五、课题年度计划	0.00%
六、课题组织实施机制及保障措施	20.23%

图9　每章内容重复率

三、相似样本列表

序号	相似比	样本编号	样本名称	申报单位	负责人	样本类型
1	3.76%	2021YF	材料与构件全流程质量[redacted]核验证	[redacted]	王[redacted]	课题
2	2.76%	2021YF	面向一体化[redacted]集成开发技术	[redacted]	徐[redacted]	课题
3	2.75%	2021YF	新型深紫外[redacted]	[redacted]	杨[redacted]	课题
4	2.75%	2021YF	大尺寸衬底[redacted]	[redacted]	胡[redacted]	课题

图 10　相似样本列表

注：为不泄露相关信息，笔者对图中部分文字进行了处理。

漏，图片、表格、公式等内容显示异常。这些问题会不同程度地影响查重结果的准确率，因此在后期需要科技管理部门对申报书的格式提出更加严格具体的规范要求，或者通过在查重系统中增加异常申报书识别功能。

（2）科技项目信息共享程度有待提高。目前，不同地区、不同层级的科技管理部门都有各自的项目计划，并且彼此之间缺乏有效的信息共享机制，这就导致查重范围局限在各自部门的内部，查重结果的可信度也大打折扣。因此，未来有必要进一步实现各级科技管理系统的互联互通，提高科技项目信息的共享程度。

（3）查重系统的智能化水平尚有很大的进步空间。设计查重系统的本意在于避免项目重复立项和检测学术不端。但应该清楚地认识到，申报书的重复比例只能作为参考，不能简单地认为重复比例高就意味着项目属于重复立项或者涉及学术不端。其中一个主要原因就是申报书中难免包含一些极易重复但影响不大的文字，如项目背景介绍、国内外研究现状等，在计算最终重复比例时它们也会被考虑在内。因此，未来可以更多地引入人工智能等技术，让系统能够“理解”出申报书的核心内容，再判断重复与否。

参考文献

[1] 王文棋，张建波，李国栋，等.国家自然科学基金申请项目人员查重方法应用和比较[J].中国科学基金，2012(5): 3.

[2] 刘如，秦潇，董晓晴，等.科技项目查重研究现状与发展对策[J].天津科技，2017, 44(2): 4.

[3] 黄思颖，蔡桂兰，徐凯，等.基于SolrCloud的分布式科技项目查重系统[J].科技管理研究，2018, 38(7): 7.

[4] 陶秀杰，周育忠，韦嵘晖，等.企业科技项目申报查重系统设计与应用[J].信息系统工程，2021(4): 3.

[5] 高爽，刘梅，屈加豹，等.智能查重方法在建设项目

环评文件技术复核中的应用探讨[J].环境影响评价, 2021, 43(6): 5.
[6] 刘玉林, 郭雅娟, 陈锦铭, 等.基于自然语言处理技术的电网招标资料查重系统研制[J].电力信息与通信技术, 2018, 16(5): 7.
[7] 侯鑫鑫, 朱文佳, 朱莉, 等.多源异构学术成果大数据的整合与揭示[J].情报理论与实践, 2021, 44(4): 162−168.
[8] 周育忠, 陶秀杰, 张自锋, 等.科技项目查重系统在企业中的实践应用[J].河南科技, 2019(28): 4.
[9] 蒋勇青, 刘芳, 于洋.学术文献相似性检测比对资源应用分析与建设策略探究: 基于万方检测系统的实证分析[J].数字图书馆论坛, 2017(12): 39−44.
[10] 吴为民.我国科研项目重复申报问题的成因与对策研究[J].农业网络信息, 2016(3): 40−43.
[11] 王春柳, 杨永辉, 邓霏, 等.文本相似度计算方法研究综述[J].情报科学, 2019, 37(3): 158−168.
[12] 奚雪峰, 周国栋.面向自然语言处理的深度学习研究[J].自动化学报, 2016, 42(10): 1445−1465.
[13] KADUPITIYA J, RANATHUNGA S, DIAS G. Sinhala short sentence similarity calculation using corpus−based and knowledge−based similarity measures[C]//Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016). Osaka, Japan: The COLING 2016 Organizing Committee. Osaka, Japan: The COLING 2016 Organizing Committee, 2016: 44−53.
[14] 来斯惟.基于神经网络的词和文档语义向量表示方法研究[D].北京: 中国科学院大学, 2016.
[15] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics. New Orleans, Louisiana: Association for Computational Linguistics. New Orleans, Louisiana: Association for Computational Linguistics, 2018: 2227−2237.
[16] 陈宏朝, 李飞, 朱新华, 等.基于路径和深度的同义词词林词语相似度计算[J].中文信息学报, 2016, 5(30): 80−88.
[17] 仲远, 王芳, 黄树成.基于百度百科多特征信息的词汇相似度计算[J].计算机与数字工程, 2020, 48(7): 1580−1584, 1736.