一种基于论文画像的科技文献数据去重算法

白文磊[1],常丽琼[1],郭 军[1,2],刘宝英[1*],甘大广[3]
(1. 西北大学 信息科学与技术学院,陕西 西安 710127;
2. 西北大学 京东人工智能与物联网联合研究院,陕西 西安 710127;
3. 万方数据有限公司,北京 100038)

摘 要:快速准确地将不同数据库中重复数据过滤清除是构建数据仓库的重要技术之一。在科技文献资源服务领域,传统的数据去重方法主要是利用数据库检索技术,进行字段内容匹配,过滤内容相同的论文数据. 然而,分布在不同数据库中的论文,一般有着不同的字段信息和字段类型,即使有相同的字段也会因为字段内容可能存在乱码信息,导致算法鲁棒性不强,这是传统搜索查找匹配方法面临的一个主要挑战。为解决这个问题,借鉴推荐系统中物品画像和人物画像算法的思想,该文提出了一种基于论文画像的科技文献数据去重算法。该算法通过 tf-idf 技术提取文章摘要中的关键字信息,再将关键字信息通过 word2vec 转换为词向量,进而计算出论文之间的相似程度并过滤掉重复数据。实验结果表明,在真实的大型论文数据集下,该算法能够有效去除重复信息,auc 均值可达到 0.98 以上。

关键词:论文画像;数据清洗;数据去重;词频-逆文档频率;词向量
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2022)08-0148-07
doi:10.3969/j.issn.1673-629X.2022.08.024

A Data Deduplication Algorithm for Scientific Literature Based on Paper Portrait

BAI Wen-lei[1],CHANG Li-qiong[1],GUO Jun[1,2],LIU Bao-ying[1*],GAN Da-guang[3]
(1. School of Information Science and Technology,Northwest Univ,Xi'an 710127,China;
2. Jingdong Joint Research Institute of AI and Internet of Things,Northwest Univ,Xi'an 710127,China;
3. Wanfang Data Co.,Beijing 100038,China)

Abstract:It is one of the important techniques for constructing data warehouse to filter and remove duplicate data from different databases quickly and accurately. In scientific literature service,the traditional data deduplication methods mainly use data searching technology to match the fields and filter out the papers with the same content. However,papers in different databases usually have different field information and field types. Even if there are the same fields,there may be garbled information in the field content,which leads to the weak robustness of the algorithm. This is a major challenge faced by traditional search and matching methods. To solve this problem,we propose a data deduplication algorithm based on the paper portrait inspired by the algorithm of item portrait and person portrait in the recommendation system. This algorithm adopts tf-idf technology to extract the keyword information in the article abstract,which are converted into word vectors by word2vec so that the similarity between papers can be calculated. The duplicate data is filtered according to their similarities. The experimental results show that the proposed algorithm can effectively filter duplicate information under the real-world data set.

Key words:paper portrait;data clean;data deduplication;tf-idf;word2vec

0 引 言

近年来,随着信息社会的全面到来和科学技术的迅猛发展,科技文献的数量呈现指数级增长,并通过互联网广泛传播[1]。为了有效利用这些科技文献资源,各种科技文献资源服务平台大量涌现,例如谷歌学术、百度文库、万方、维普、IEEE、ACM、CNKI 等。相比传统的纸质文献资料,这些基于云计算和虚拟化技术构建的文献资源库[2-3],可以让人们通过网络快速检索

收稿日期:2021-04-19 **修回日期**:2021-08-24
基金项目:国家重点研发计划项目(2017YFB1400301)
作者简介:白文磊(1997-),男,硕士生,研究方向为机器学习与推荐系统;通信作者:刘宝英(1966-),女,博士,教授,研究方向为文物保护与人工智能。

出需要的文献资料,为科学研究和技术开发提供便利。但是,由于科技文献资源的服务比较分散,各种科技文献资源数据库相互独立,造成科技文献资源的共享和协同服务能力比较弱。当查找多个数据库中的文献时,人们需要分别登录访问不同数据库,然后人工进行筛选鉴别感兴趣的信息。显然,这种方式非常繁琐,需要消耗人们较多的时间和精力,为了方便科技人员查找分布在不同数据库中的文献资料,构建一个统一的数据仓库是解决问题的有效途径。

在构建数据仓库的过程中,文献数据的去重[4]处理是一个非常关键的问题。因为不同的数据库往往包含相同的文献数据,由于存储格式和文件结构的不同,这些相同的文件不易区分,造成数据冗余重复,而大量的冗余数据会消耗存储资源,降低访问效率。针对这一问题,基于论文画像技术,提出一种科技文献数据去重算法。该算法将论文关键字转换为词向量,通过计算论文之间的相似程度,过滤掉重复数据。该算法在实验数据集和真实数据集上都取得了比较满意的结果。

主要贡献包括 4 个方面:

(1)将推荐系统中的物品画像算法引用到文献数据去重的工作中,提出了一种基于论文画像的数据去重算法,通过分析论文关键字信息,得到存储在不同数据库中的重复数据;

(2)将自然语言处理中的词频-逆文档频率技术引用到了论文主要内容提取的任务中,避免了传统匹配算法因为个别文献的乱码问题导致算法精确率、召回率降低的问题;

(3)通过使用 word2vec,将词频-逆文档频率方法提取的关键字信息转化为向量,从而更好地计算论文之间的相似程度;

(4)不同数据集的应用实验结果表明,该算法能够准确完整地检测出重复的论文,并且与已有的方法相比,有更好的鲁棒性,且不降低算法的精确率与召回率。

采用的关键数学符号如表 1 所示。

表 1　采用的关键数学符号

符号	意义
m	数据集中文章数
k	关键词个数
p	关键词向量维度
word_vec	关键词向量
content_vec	文章内容向量
W'	相似度阈值
V	词袋中单词种类数量
N	隐含层神经单元个数

1　相关工作

目前,数据去重技术广泛应用于数据存储、备份和归档系统[5-6]。数据去重技术主要分为相同数据检测技术与相似数据检测。相同数据检测主要包括相同文件及相同数据块两个层次。相似数据检测利用数据自身的相似性特点,通过 shingle 技术[7]、bloom filter 技术[8]和模式匹配算法[9]挖掘出相同数据检测技术不能识别的重复数据。

完全文件检测技术(whole file detection,WFD):WFD 技术[10]以文件为粒度查找重复数据。WFD 首先通过对整个文件进行 hash 计算,然后与数据库中以存储的 hash 值进行比较,如果检测到相同的值就删除重复数据,否则存入数据库中。该方法执行效率高,可以检测到所有完全相同的文件,但不能检测不同文件内部的相同数据。

针对 WFD 算法的缺陷,有研究者提出了细粒度块级别的去重算法—固定块检测技术(fix-sized partition,FSP)[11],该算法根据预先定义好的一个块大小,将所有文件按照这个块大小进行划分,然后将每个数据块进行 hash 计算,与数据库中已有的 hash 值进行比较,相同删除,不同存入。该算法较好地解决了 WFD 算法问题。但是,该算法也有一定的局限性,不能根据文件内容和文件之间的关系进行调整和优化,例如:对于插入问题(在原始数据流中插入少量的新字节)和删除问题(在原始数据流中删除少量字节)处理效率比较低。

另一方面,互联网技术和电子商务的迅速发展促进了推荐系统的发展成熟[12]。其中基于标签的推荐算法[13](tag-based,TB)在推荐系统中得到了广泛应用。用户用标签来描述对物品的看法,因此标签是联系用户和物品的纽带,也是反映用户兴趣的重要数据源。该算法的思想如下:统计每个用户最常用的标签;对每个标签,统计被打过这个标签次数最多的物品;对于用户,首先找到他最常用的标签,然后将具有这些标签的最热门的物品推荐给用户。该算法较好地利用了用户对物品的作用信息进行推荐,但需要用户对物品主动打标签。为此,又有学者提出了从物品本身信息中提取出关键字[14-15]构建出物品画像,将关键字信息反作用到用户的基于内容的推荐算法(content-based,CB),从而解决了 TB 算法的缺陷。受 CB 算法启发,该文提出了一种基于论文画像的科技文献数据去重算法。

2　基于论文画像的科技文献数据去重算法

本节主要介绍词频-逆文档频率、词向量以及相似度计算方法,并将其引入到论文画像、数据去重的工

作中。

2.1 词频-逆文档频率

词频-逆文档频率(term frequency - inverse document frequency,tf-idf)技术[15],是一种用于信息检索、关键字提取的常用加权技术。一个词语的重要性随着它在该文档中出现次数呈正比,同时也随着它在语料库中出现的文档数呈反比。例如,某个词语在其他论文中很少见,但在该论文中多次出现,那么它很有可能反映了这篇论文的主题,即该词语就可以认为是关键词。

tf-idf 技术原理如下:

以统计一篇论文中的关键词为例,要想得到该论文的主题,最简单的方法就是统计该论文中每个词出现的次数占论文总词数的百分比,即词频(term frequency,tf),其计算公式如下:

$$\mathrm{tf} = \frac{|\mathrm{word}|}{|\mathrm{words}|} \tag{1}$$

其中,word 表示需要统计词频的单词,|word| 表示该单词在论文中出现的次数,|words| 表示论文单词总词数。

通过 tf 方法计算出的频率最高的几个词也就是这篇文章的关键词。但仅仅使用 tf 方法,出现频率最高的词很容易是停顿词,例如:of,on,in 等,这些词很明显无法反映论文的主要内容,所以还需要给这些停顿词添加一个惩罚,这个惩罚就是逆文档频率 idf。

逆文档频率度量一个单词的普遍重要性,它的大小与该词的常见程度呈反比,其计算方法是语料库中的论文总数除以语料库中包含该词语的论文数,再取对数,计算公式如下:

$$\mathrm{idf} = \log \frac{|\ \mathrm{papers}\ |}{|\ \mathrm{content\_papers}\ | + 1} \tag{2}$$

其中,|papers| 表示论文总数,|content_papers| 表示包含某个关键字的论文数量。

在得到 tf,idf 计算公式后,就可以计算 tf-idf 的值,计算公式如下:

$$\mathrm{tf-idf} = \mathrm{tf} \times \mathrm{idf} \tag{3}$$

由此,可以解决上述停顿词占比很高的问题。一般地,停顿词会在每篇文章中出现,导致 idf 接近于 0,从而停顿词的 tf-idf 值也很低。

2.2 word2vec 编码

在通过 tf-idf 技术提取到每篇论文的关键词后,将这些关键词交给机器学习算法处理,但机器无法理解这些语言,因此首先要做的就是将这些关键词转换为编码。一种最常见的编码方式为 one-hot representation 编码:假设词袋中共有 V 种单词,则设置一个 V 维的向量,向量的分量中只有一个为 1,其余全为 0,1 的位置对应该词在词袋中的索引。但这种编码方式在文本特征表示中存在以下缺点:

(1)在文本中词的顺序信息是很重要的信息,one-hot 编码是一个词袋模型,没有考虑文本中词与词之间的顺序。

(2)one-hot 编码丢失了词与词之间的关系信息,无法体现单词与单词的关系的远近程度。

(3)每个单词的 one-hot 的编码维度是整个词袋中单词种类的数量,容易造成维度灾难。

Distributed representation 可以解决 one-hot 编码的上述问题。它的思路是通过一个简单的神经网络模型,将每个词都映射到一个较短的词向量上,该词向量的维度需要在训练时自己指定。

word2vec 模型[16]就是一种三层的神经网络,输入的是单词的 one-hot 编码,隐含层的激活函数是线性激活函数,输出层的维度与输入层的维度相同。这个模型类似于自编码器的网络模型,但与自编码器不同的是,当这个模型训练好以后,并不会用这个模型去处理,预测新的输入,而是提取网络模型中隐含层的权重。

word2vec 模型一般分为两种,分别是:CBOW 与 Skip-Gram 模型[16]。其中 CBOW 模型是将一个词的上下文作为网络的输入,再预测这个词,其网络结构如图 1 所示。

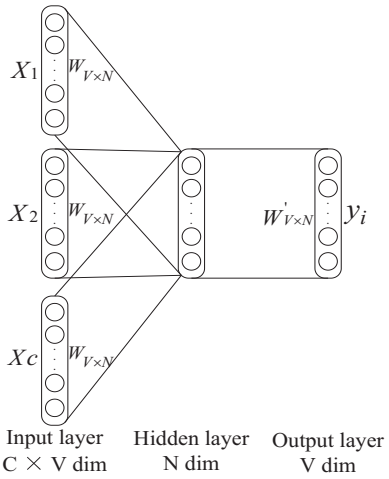


图 1 CBOW 网络结构示意图

以 CBOW 模型为例,介绍该模型的训练过程:

(1)输入层为上下文单词的 one-hot 编码,假设单词向量的维度为 V,共有 C 个单词。

(2)所有的 one-hot 编码分别点乘共享矩阵 $\boldsymbol{W}_{V\times N}$,$N$ 为隐含层神经单元个数,同时也是映射后的单词短向量的维度,一般地,V 远小于 N。

(3)将步骤(2)中得到的 C 个向量,相加后求平均作为隐含层向量 size=(1,N)。

(4)将隐含层向量与输出层权重矩阵 $\boldsymbol{W}'_{V\times N}$ 点乘,

并经过输出层的 softmax 函数处理得到 V 维的概率分布。

(5)概率最大的 index 所指示的单词即为预测的中间词与真实标签中 one-hot 编码作比较,误差越小越好。

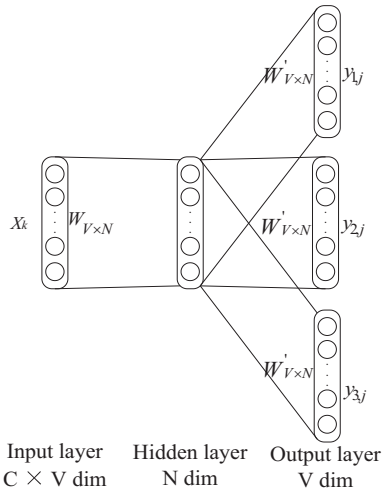


图 2　Skip-Gram 网络结构示意图

这里,需要注意的是 word2vec 的输出层经过 softmax 激活函数得到一组概率分布,这和机器学习中的多分类问题相似,故选取交叉熵损失函数作为代价函数,交叉熵损失函数的定义如下:

$$L = \frac{1}{M}\sum_{i} - \sum_{c=1}^{V} y_{ic}\log(p_{ic}) \tag{4}$$

其中, M 为样本数量, y_{ic} 为指示变量, p_{ic} 为观测样本属于类别 c 的预测概率。

通过梯度下降法就可以更新隐含层与输出层的权重矩阵 $\boldsymbol{W}$ 和 $\boldsymbol{W}'$。结束训练后,若想要得到某个单词的词向量,则用该向量的 one-hot 编码与 $\boldsymbol{W}$ 矩阵点乘,所得到的结果便是该词的词向量。

2.3　相似度计算方法

通过词频-逆文档频率,按照每个词的权重从高到低排序,截取前 k 个关键字{keyword_1,keyword_2,…,keyword_k},然后通过 word2vec 方法,将每篇论文的关键字转换为关键字向量,再取关键字向量和的加权平均值(如式(5)),就可以得到每篇论文的文章特征向量。

假设关键词向量的维度为 p,则 k 个关键字就组成了一个 $k \times p$ 的矩阵:

$$\text{content\_vec} = \frac{\sum_{i}^{k} \text{word\_vec}_i}{|k|} \tag{5}$$

计算向量之间相似度的方法主要有以下几种:余弦相似度、欧氏距离、曼哈顿距离、皮尔逊相关系数[17]等,具体计算公式如表 2 所示。该文使用余弦相似度计算论文之间的相似度。

表 2　相似度计算方法

方法	描述
余弦相似度	$\text{sim}(A,B) = \dfrac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2 \times \sum_{i=1}^{n}(B_i)^2}}$
欧氏距离	$d(A,B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$
曼哈顿距离	$d(A,B) = \sum_{i=1}^{n}\lvert A_i - B_i\rvert$
皮尔逊相关系数	$\text{sim}(A,B) = \dfrac{\sum_{i=1}^{n}(A_i-\bar{A})(B_i-\bar{B})}{\sqrt{\sum_{i=1}^{n}(A_i-\bar{A})^2}\sqrt{\sum_{i=1}^{n}(B_i-\bar{B})^2}}$

最后,该算法可分为如下步骤:

(1)分别加载含有不同字段的论文数据信息。

(2)将各个数据信息提取公共字段并组合为一个数据集。

(3)使用 tf-idf 方法提取文章前 TOP-K 个关键字信息。

(4)使用合并后的数据集,训练 word2vec 模型。

(5)使用 word2vec 将关键字信息转换为词向量。

(6)将 k 个关键字求和再去平均得到每篇文章的文章向量。

(7)根据提取到的文章向量,以及表 2 中的向量相似度计算方法,计算任意两样本之间的相似度,生成 $m \times m$ 维的对称矩阵。

算法将相似度值高于阈值 λ 的论文对提取,需要注意的是:为避免将重复的论文对放入候选集中,只对对称矩阵的上三角进行遍历,同时也降低了算法的时间复杂度。

3　实验与结果

在本节中,将详细描述实验中所用到的数据集、超参数确定、算法度量指标以及实验结果和分析。

3.1　数据集

在 CiteULike Dataset[18]、Citation Network Dataset[19]、Covoid-19 Paper Dataset[20]以及 Arxiv Paper Dataset[21]这 4 个数据集上验证提出的算法的有效性。

为避免人工标注,产生标签等繁琐步骤,将上述数据集分别随机抽取 20% 的样本作为重复论文重新插入数据中,并打乱顺序。

3.2　度量指标

将数据的去重工作归为机器学习中的二分类问

题,即:判断某一篇论文是否和数据库中已有论文相似。对于二分类问题,常用的指标有:精确率、召回率、Precision and Recall (P-R) 曲线、Receiver Operating Characteristic (ROC)曲线[22]等。

精确率:是指分类正确的正样本个数占分类器判定为正样本的样本个数的比例。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

召回率:是指分类正确的正样本个数占真正的正样本个数的比例。

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

为综合评估算法的质量,采用绘制 P-R 曲线、ROC 曲线的方法来验证算法的有效性。

P-R 曲线:横轴为召回率,纵轴为精确率。对于一个二分类模型,P-R 曲线上的一个点代表着,在某一阈值下,模型将大于该阈值的结果判定为正样本,小于判定为负样本,此时返回结果对应的召回率和精确率。原点附近代表着当阈值最大时模型的精确率与召回率。

ROC 曲线:横坐标为假阳性率(false positive rate, FPR),纵坐标为真阳性率(true positive rate,TPR)。算法应当使得 FPR 尽可能的小而 TPR 尽可能的大,FRR 与 TPR 计算方法如下:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{8}$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

3.3 实验结果及分析

在上述四个数据集以及 CBOW 模型上进行了三组实验,分别如下:

实验一:相似度阈值 λ 对算法的影响。

精确率和召回率是既矛盾又统一的两个指标,为了提高精确率,算法需要尽量在“更有把握”时,即采用更高的相似度阈值 λ 把样本预测为重复样本,但此时往往会因为 λ 过高而漏掉许多“没有把握”的重复样本,导致回归率很低。为综合评估算法的性能,在四种数据集上绘制了 P-R 曲线以及 ROC 曲线,如图 3、图 4 所示。在该组实验中设置 $k = 10$, $N = 100$。

图 4 图例中 auc 代表 ROC 曲线下的面积大小,auc 越大,模型性能越好。

观察图 3 可知,当召回率接近 0 时,模型在 4 个数据集的精确率都是 0.9 以上。并且随着召回率的增加,精确率整体呈下降趋势,这与之前的分析相吻合。并且分析图 4 曲线下的面积,即 auc 值均在 0.98 以上,取得了较为不错的实验结果。

实验二:关键词个数 k 对算法的影响。在该组实验中,设置 $N = 100$, k 的取值为 5、10、20、40、60。

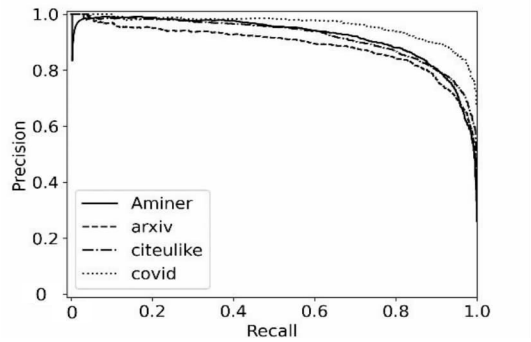


图 3 P-R 曲线

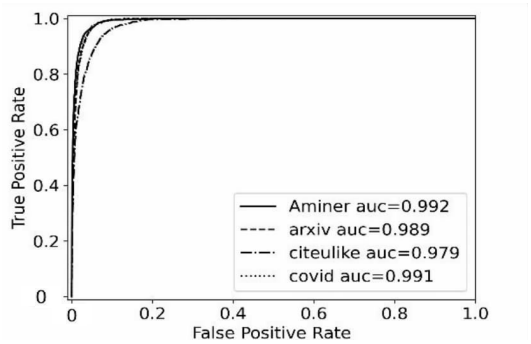
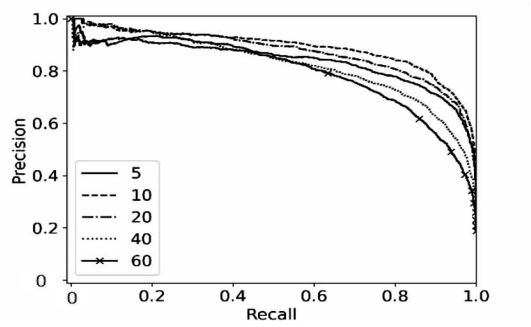
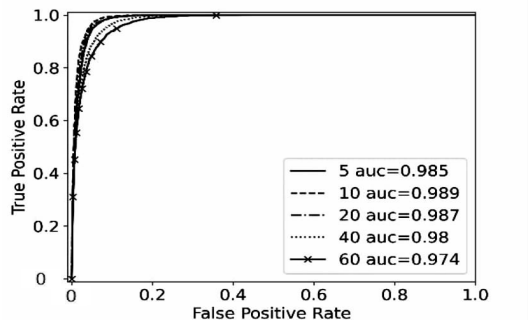


图 4 ROC 曲线

关键词个数 k 对算法的实验结果有着重要影响。k 值过小,可能导致将不重复的样本归为重复样本,导致精确度降低。由于算法是将 k 个关键字的词向量求和后平均作为样本的画像,k 值过大,会导致各个样本的画像趋于平均化。为此,在 Arxiv Dataset 和 CiteUlike Dataset 上绘制了 P-R 曲线以及 ROC 曲线,如图 5、图 6 所示。

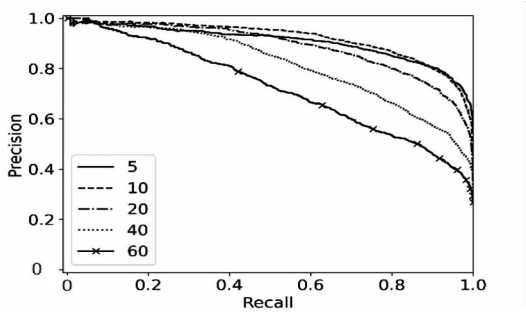


(a)

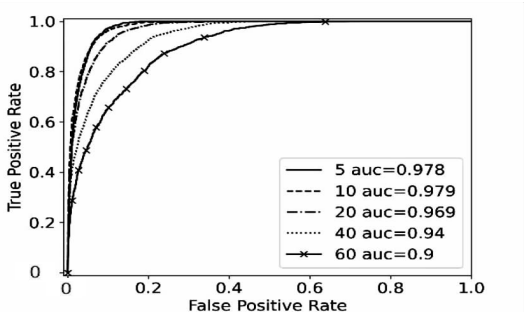


(b)

图 5 Arxiv 数据集下的 P-R 曲线(a)和 ROC 曲线(b)



(a)



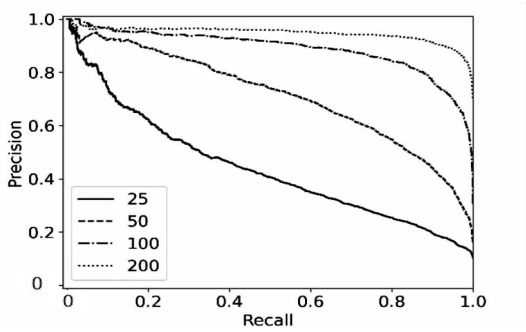
(b)

图 6　CiteUlike 数据集下的 P-R 曲线(a)和 ROC 曲线(b)

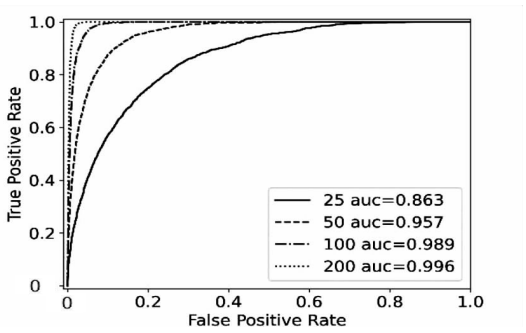
观察图 5、图 6 可知,当 k 取 10 时,算法的效果达到最优。

实验三:隐含层神经元数量 N 对算法的影响。

在该组实验中,设置 k = 10, N 的取值为 25,50,100,200。



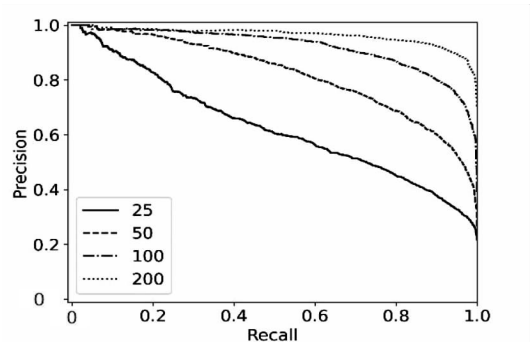
(a)



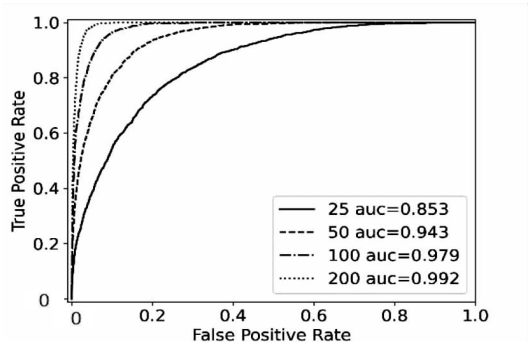
(b)

图 7　Arxiv 数据集下的 P-R 曲线(a)和 ROC 曲线(b)

图 7、图 8 中 P-R 曲线与 ROC 曲线在不同的隐含层神经元数据 N 上实验效果差距很大,且 N 值越大模型效果越好。这是因为 N 值过小,会导致论文画像不精准、相似度计算不准确,从而降低算法的精确率与召回率。



(a)



(b)

图 8　CiteUlike 数据集下的 P-R 曲线(a)和 ROC 曲线(b)

4　结束语

将推荐系统中的人物、物品画像算法,应用于文献资源数据仓库构建中的数据去重工作,提出了一种基于论文画像的科技文献数据去重算法。通过使用 tf-idf,word2vec 技术提取到文章向量,进而计算出论文之间的相似程度。通过去除相似程度高于阈值的论文,完成数据去重. 实验结果表明,该算法能够准确地检测出数据库中重复的论文,实现有效多数据去重。在参数调优后,auc 均值可达到 0.98 以上。该算法在万方文献数据集上进行了初步的应用测试,取得了不错的效果。

参考文献:

[1] MAYER-SCHÖNBERGER V,CUKIER K. Big data:a revolution that will transform how we live,work,and think[M]. New York:Houghton Mifflin Harcourt,2013.

[2] FOX A,GRIFFITH R,JOSEPH A,et al. Above the clouds:a berkeley view of cloud computing[R]. California:University of California,2009.

[3] UHLIG R,NEIGER G,RODGERS D,et al. Intel virtualiza-

tion technology[J]. Computer,2005,38(5):48-56.
[4] STORER M W,GREENAN K,LONG D D E,et al. Secure data deduplication[C]//Proceedings of the 4th ACM international workshop on storage security and survivability. New York:ACM,2008:1-10.
[5] 敖 莉,舒继武,李明强. 重复数据删除技术[J]. 软件学报,2010,21(5):916-929.
[6] 张兴兰,何丹丹. 基于改进的 Simhash 算法的相似文档识别技术[J]. 计算机科学与应用,2020,10(2):371-378.
[7] AGHAYEV A,TS'O T,GIBSON G,et al. Evolving ext4 for shingled disks[C]//15th USENIX conference on file and storage technologies. Santa Clara:USENIX,2017:105-119.
[8] PATGIRI R,NAYAK S,BORGOHAIN S K. rDBF:a r-dimensional bloom filter for massive scale membership query [J]. Journal of Network and Computer Applications,2019,136:100-113.
[9] GAL A, ROITMAN H, SHRAGA R. Learning to rerank schema matches[J]. IEEE Transactions on Knowledge and Data Engineering,2019,33(8):3104-3116.
[10] CLEMENTS A,AHMAD I,JINYUAN L I,et al. Computer storage deduplication:U. S. ,10,642,794[P]. 2020-05-05.
[11] BOBBARJUNG D R,JAGANNATHAN S,DUBNICKI C. Improving duplicate elimination in storage systems[J]. ACM Transactions on Storage,2006,2(4):424-448.
[12] BAWDEN D,ROBINSON L. Information overload: an introduction[M]//Oxford research encyclopedia of politics. [s. l.]:[s. n.],2020.
[13] VIG J,SEN S,RIEDL J. Tagsplanations:explaining recommendations using tags[C]//Proceedings of the 14th international conference on intelligent user interfaces. New York: ACM,2009:47-56.
[14] PAZZANI M J,BILLSUS D. Content-based recommendation systems [M]//The adaptive web. Berlin: Springer, 2007:325-341.
[15] RAMOS J. Using TF-IDF to determine word relevance in document queries[C]//Proceedings of the first instructional conference on machine learning. [s. l.]:[s. n.],2003:133-142.
[16] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[J]. arXiv: 1301.3781,2013.
[17] SARWAR B,KARYPIS G,KONSTAN J,et al. Item-based collaborative filtering recommendation algorithms[C]//Proceedings of the 10th international conference on World Wide Web. New York:ACM,2001:285-295.
[18] WANG H,CHEN B,LI W J. Collaborative topic regression with social regularization for tag recommendation [C]// Twenty-Third international joint conference on artificial intelligence. Beijing:AAAI Press,2013.
[19] TANG J,ZHANG J,YAO L,et al. Arnetminer: extraction and mining of academic social networks[C]//Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2008:990-998.
[20] WANG L,LO K,CHANDRASEKHAR Y,et al. CORD-19: the Covid-19 open research dataset[J]. arXiv:2004.10706, 2020.
[21] CLEMENT C B,BIERBAUM M,O'KEEFFE K P,et al. On the use of ArXiv as a dataset[J]. arXiv:1905.00075,2019.
[22] BRADLEY A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms[J]. Pattern Recognition,1997,30(7):1145-1159.