计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(6):149−157 [doi: 10.15888/j.cnki.csa.008497]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于文本分析的标书综合评估模型①

任　杰

(中国水利电力物资集团有限公司, 北京 100043)
通信作者: 任　杰, E-mail: renjie@cweme.com

摘　要: 人工智能的发展为传统的标书评估方法的优化和改进提供了新的方法, 针对标书评估中人工评标效率低和难以有效识别围标、串标行为的问题, 本文提出了一种基于文本分析的标书综合评估模型, 包含文本评估模型和文本评级模型, 模型为更客观、科学、智能化地进行工程建设项目的评标和防范围标、串标提供支持. 首先, 构建文本评估模型, 在传统的评标指标中加入基于 Shingling 算法计算的投标文件的重复率, 并将投标文件要求的模板目录与真实目录进行匹配对比计算投标文件的响应程度, 利用层次分析法计算文本评估指标的权重. 然后, 构建文本评级模型, 基于权重改进的 Simhash 算法计算投标文件相似度, 采用企业关联度、资质与报价的一致性、价格上(下)浮率、企业关联度、异常行为等评级指标, 通过综合评级获得投标文件的评级结果, 协助评标专家识别围标、串标行为. 最后, 通过文本评估模型定量计算得到标书得分排序, 通过文本评级模型定性分析得到标书识别围标、串标的结果, 两者共同实现了对标书的综合评估.

关键词: 标书评估; 文本评估; 文本评级; 相似度; 重复率; 文本分析

引用格式: 任杰.基于文本分析的标书综合评估模型.计算机系统应用,2022,31(6):149–157. http://www.c-s-a.org.cn/1003-3254/8497.html

Comprehensive Evaluation Model of Bidding Documents Based on Text Analysis

REN Jie

(China National Water Resources & Electric Power Materials & Equipment Co. Ltd., Beijing 100043, China)

Abstract: The development of artificial intelligence provides a new tool for the optimization and improvement of traditional bid evaluation methods. To tackle the problems of low manual bid evaluation efficiency and difficulty in identifying bidding collusion, this study proposes a comprehensive evaluation model for bid documents based on text analysis, including a text evaluation model and a text rating model. The proposed model can provide support for a more objective, scientific, and intelligent bid evaluation of construction projects and the prevention of bidding collusion. First, a text evaluation model is built. The repetition rate of bid documents calculated by the Shingling algorithm is added to traditional bid evaluation indicators. Then, the template catalog required by the bid documents is compared with the real catalog for the calculation of the response level of the bid documents. The analytic hierarchy process is employed to determine the weight of the text evaluation index. Next, a text rating model is built. The similarity between bid documents is calculated by the weight-improved Simhash algorithm. The comprehensive rating is performed with corporate relevance, consistency between qualifications and quotations, price floating, abnormal behavior, etc. as rating indicators. The rating results of bid documents are helpful for bid evaluation experts to identify bidding collusion. Finally, a quantitative calculation is conducted with the text evaluation model to yield the bid score ranking, and a qualitative analysis is made with the text rating model to present the results of identifying bidding collusion. Taken together, the two realize the comprehensive evaluation of bid documents.

Key words: bid evaluation; text evaluation; text rating; similarity; repetition rate; text analysis

① 收稿时间: 2021-08-11; 修改时间: 2021-09-13; 采用时间: 2021-09-29; csa 在线出版时间: 2022-03-11

近年来, 我国的社会发展与经济建设取得了举世瞩目的成绩. 社会的发展过程离不开工程建设, 工程建设招标和投标是在市场经济条件下进行工程建设的一种经济活动, 其实质是一种市场竞争行为. 在甲方市场的条件下, 招标人可以通过招标活动在众多投标人中选定报价合理、工期较短、信誉良好的承包商、供应商来承担工程建设任务[1]. 工程建设的招投标不仅具有高报价、高复杂性和高竞争性等问题, 还存在人工评标效率低和识别围标、串标行为难的问题[2,3]. 这些问题都在不同程度上阻碍了工程的建设和企业的发展, 同时也给招标投标的工作带来了不小的挑战. 因此招标投标的各个环节是否能够遵守高效、客观、科学、公平、公正、公开的原则至关重要[4].

目前招投标领域正在由纸质化招标向电子化招标的方向发展, 这也为利用计算机分析电子化招投标文件提供了可能. 首先, 利用计算机对标书进行评估, 可以实现对标书的预选, 为人工评分提供了参考和客观依据; 其次, 计算机的应用与分析为构建电子化招投标系统和标书文本分析工作提供了条件; 最后, 利用计算机分析招投标过程信息和背景信息, 可以为识别围标、串标行为提供参考. 但是目前招投标实践中, 标书评估主要还是依靠人工评标, 缺少全面、科学的技术辅助手段. 招投标研究领域中, 利用大数据分析标书并识别围标、串标的技术仍然不完善, 缺乏通用性. 这主要是因为投标过程具有高复杂性, 现有的方法仅仅针对一个或两个指标进行定量分析, 这显然是不够的. 标书文本的分析不仅要考虑内部、外部等多个指标, 还需将定量分析与定性分析相结合, 从而实现更加全面、完整、科学的标书评估.

随着深度学习在 NLP 领域的发展, 利用 NLP 进行自然语言理解 (natural language understanding, NLU) 和自然语言生成 (natural language generation, NLG) 已经越来越普遍[5]. 文本是语言信息的主要载体, 利用文本信息进行挖掘并提取关键信息, 对于人们快速准确地获取文本内容具有重要的作用. 语义相似度计算 (semantic textual similarity) 是联系文本信息表示和潜在上层应用之间的纽带[6], 重复率常用于大型网页和巨量文本的量化计算[7,8]. 在相似度和重复率的实践上, 目前 Simahash 算法和 Shingling 算法[9] 被认为是当前最好的算法之一[10,11]. 采用这两种算法计算投标文件间的相似度与重复率, 可以为标书文本的评估和识别围标、串标行为提供量化指标.

本文提出了基于文本分析的标书评估模型, 从定量分析和定性分析两个方面分别处理标书文本, 实现对标书的综合评估. 本文第 1 节介绍评估模型的框架和基本思路, 第 2 节介绍涉及到的关键算法与改进, 第 3 节介绍模型的评估指标及计算方法, 第 4 节进行实际案例分析, 第 5 节总结评估模型, 提出不足与展望.

1 评估模型框架

当前招投标研究领域主要存在两个主要问题: (1) 识别围标、串标行为主要依赖评标现场进行人工识别和判断, 但是评标现场时间有限, 并且围标和串标行为往往不易发现, 缺少有效的机器辅助手段; (2) 当前评标工作中, 利用计算机分析标书时缺少有效合理的评价指标和评价方法, 现有评价指标往往侧重于对少数几个方面进行定量分析, 缺少结合定量分析与定性分析的全面评价体系.

本文提出了基于文本分析的标书综合评估模型, 模型通过基于定量分析的文本评估和基于定性分析的文本评级实现对标书的综合评估. 文本评估模型是通过定量分析计算 5 项指标及权重得到标书评分, 通过评分对标书进行排序, 为实际评标工作中的标书评分提供参考. 文本评级模型是通过定性分析利用 7 项指标分别对标书文本进行评级得到评级结果, 通过评级结果识别投标企业是否疑似出现围标、串标行为, 模型识别再结合人工核查确认最终的识别结果, 模型为评标工作中识别围标、串标行为提供参考. 文本评估和文本评级的结果分别实现了对标书的定量计算和定性分析, 两者结果综合集成后即可实现对标书的综合评估, 标书评估模型框架图见图 1.

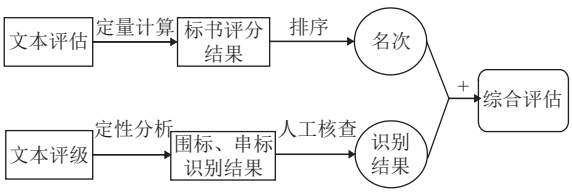

图 1 标书评估模型框架图

本文的研究语料来源于中国水利电力物资集团有限公司工程建设中的招投标文件, 投标文件通常包含投标函部分、商务标部分和技术标部分. 由于投标文件是非结构化文本数据, 而且文件中不仅含有大量的文字信息, 还有表格和图片信息. 这些非结构化信息给

开标现场的评标带来了不少困难, 尤其是投标企业出现围标、串标行为时, 评标专家难以在评标现场短时间内进行定量或定性识别. 《中华人民共和国招投标法》《招投标实施条例》《招投标实施细则》等法律法规规定了对出现围标、串标行为的处罚条例, 但是缺少于围标、串标行为的界定标准.

在文本分析方面, 构建招投标过程文件查重对比模型, 通过基于 NLP 的权重改进的 Simhash 算法和 Shingling 算法对投标文件进行分析, 得到投标文件之间的相似度和重复率. 再通过匹配和对比得到招标文件目录的匹配度、资质与报价的一致性和投标价格的上 (下) 浮率指标. 这些指标通过定量分析为评标专家的评标工作提供更加客观、准确、科学的依据, 同时也为识别围标、串标的行为提供了参考.

在行为分析方面, 构建异常检测模型, 针对投标企业在投标过程中出现的异常行为进行分析, 从而识别企业是否存在围标、串标的嫌疑. 其中异常行为包括: 故意废标、开标前几家企业同时撤回标书、不同企业的保证金出自同一账户、投标文件签名字迹一致、标书出现明显的错误等.

在背景分析方面, 构建企业资质审查模型, 首先建立基于知识图谱的文本知识库, 实现知识的智能存储、智能关联、智能推理, 通过企业与项目之间的关系, 形成网状的知识结构, 利用知识问答、实体查询、关系查询、逻辑推理等功能, 实现对企业关联度的分析计算. 然后利用基于 OCR 技术的企业资质审查模型, 对投标企业资质进行审查, 通过 OCR 识别自动抽取投标文件中的企业资质等证书图片信息, 获取证书的名称、编号和印章信息, 将证书名称和编号上传至查验网站进行真伪查验, 再对印章信息进行真实性查验, 确定证书的真实性和有效性. 然后利用政府的公开信息查询企业是否出现违规、失信等情况, 得到企业的信用度.

基于文本分析的标书综合评估模型在传统的评估指标上加入文本方面、行为方面和背景方面的综合分析, 构成了更加全面、客观的标书综合评估模型, 模型的评估指标框架图见图 2.

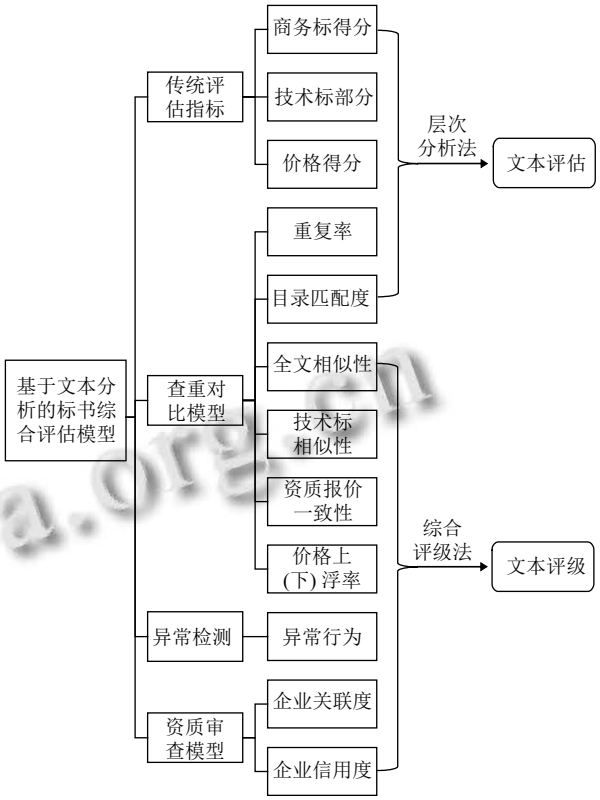

图 2 评估指标框架图

2 算法介绍与改进

标书文本分析的核心技术为文本相似度和重复率计算, 文本相似度是定性分析两个文本是否具有相似性, 文本重复率是定量计算两个文本的重复程度.

2.1 改进的 Simhash 算法

传统的文本相似度是通过计算文本特征词所构成的特征向量的夹角余弦值实现的, 面对长文本, 传统的方法由于整个特征向量的维度高, 导致计算的时间和空间复杂度都很高. 面对几万字的标书, 传统的相似度计算方法效率过低.

Simhash 算法解决了无法处理长文本的问题, 并常常被用于实践, Simhash 是一种局部敏感哈希, 局部敏感是指假如两个字符串具有一定的相似性, 这种相似性在哈希之后仍然会被保持, 这种特性常用于海量文本之间的相似度计算, 最早被 Google 应用于对海量文本进行去重处理[12]. Simhash 是一种降维的思想, 它将高维的向量映射成低维的向量并得到一个 Simhash 值, 即一个n位的指纹, 而相似文档的指纹之间只存在少量的不同, 因此通过计算n位指纹的海明距离即可判断文本之间的相似度[13,14].

Simhash 算法是由 Manku、Jain、Sarma 3 位 Google 工程师提出并通过实验验证了采用 64 位的指纹时, 文本间的海明距离取$k = 3$作为阈值来判断文本的相似是合理的. 由于参数k的取值直接影响算法的准

确率和召回率, 这两个指标大致呈现反比关系, 实验发现当$k=3$时, 算法的准确率和召回率均在 75% 左右, 并且达到了较好的均衡[15], 适用于标书文本的相似度计算. 除此之外, Simhash 算法通过降维的思想将高维特征向量映射成唯一的二值 Simhash 值, 降低了计算复杂度, 提升了算法效率.

传统的 Simhash 算法在权重计算时通常直接设置为 1 或者特征词的词频, 这就无法体现出词汇的分布特征, 导致信息的丢失和准确率降低. 为了解决传统 Simhash 算法中权重计算不充分的问题, 受文献 [16] 的启发, 本文在权重计算中使用词频-逆向文件频率 (TF-IDF) 和信息熵的基础上, 加入了特征词偏向性权重, 并人为判断特征项是否能够作为算法特征项进行计算, 最终形成了基于熵-特征词偏向性加权的 Simhash 算法, 具体计算方法如下.

(1) 词频-逆向文件频率定义为:

$$w_k = tf(t_k,d_j)\times idf(t_k) \tag{1}$$

其中, $tf(t_k,d_j)$代表词频, 是指特征项t_k在文本d_j中的词频, $idf(t_k)$代表逆向文件频率, 是指语料库中文件总数与出现特征词t_k的文件数量的比值的对数.

(2) 左右信息熵和熵量分别定义为:

$$H_l(w) = -\sum_{\forall a\in A} P(aw|w)\times lbP(aw|w) \tag{2}$$

$$H_r(w) = -\sum_{\forall b\in A} P(wb|w)\times lbP(wb|w) \tag{3}$$

$$H_k(w) = \frac{H_l(w)+H_r(w)}{2} \tag{4}$$

其中, w为单词, $H_l(w)$为单词的左熵, $P(aw|w)$为单词左侧出现不同词的频率, a表示与w结合的词. $H_r(w)$为单词的右熵. $H_k(w)$为熵量.

(3) 特征词偏向性权重定义为:

$$E_k = \max(a_i) \tag{5}$$

其中, a_i是特征项所属的标书部分 (标书通常分为: 投标函部分、商务标部分、技术标部分) 的权重, 该权重是通过对评标专家对各部分重要性排序通过层次分析法计算获得.

(4) 基于熵-特征词偏向性加权公式:

$$W(t_k,d_j) = \sqrt{\frac{(w_k)^2+(H_k)^2+(E_k)^2}{3}} \tag{6}$$

上述公式的物理意义是: 特征项t_k在文档d_j中出现次数越多, 在所有文档中出现次数越少, 信息量越大, 所属标书部分重要性程度越高, 则其对应的权重越大.

(5) 特征项的二次选择

经过上述步骤计算出来的特征项及对应的权重在带入 Simhash 算法进行计算之前, 需要结合标书文本的特殊性和本次投标所属行业关键信息的专业性利用预定的阈值进行人工二次选择, 通过二次选择提高特征项的准确性和代表性, 从而提高 Simhash 算法的计算效果.

(6) Simhash 值和海明距离的计算

Simhash 算法主要有 2 个主要步骤: 计算 simhash 值和计算文本间的海明距离.

1) 计算 Simhash 值.

首先, 对于给定的标书文本, 利用停用词表过滤掉符号、助词、语气词等无效字符, 然后通过分词库进行分词, 将文本转换为一些特征词的集合$(a_1,a_2,\cdots,a_n)$, 集合中各元素的权重$(w_1,w_2,\cdots,w_n)$为该特征词在文本中的词频. 然后, 通过 hash 计算将集合中每个特征词映射为长度为n的二进制数 hash 值[17], 再将二进制数中的 0 变为−1, 并乘以权重. 最后把乘以权重后的特征集合按位累加, 得到一个n位的文本特征值 (即文本的指纹). 遍历文本特征值的每一位, 当该位值大于 0 时赋值为 1, 小于等于 0 时赋值为 0, 即可得到降维后的文本的 Simhash 值, 算法流程图见图 3.

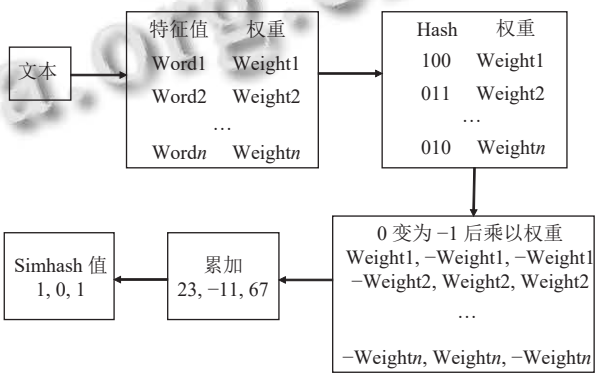

图 3 Simhash 算法流程图

2) 计算海明距离 (Hamming distance).

由于相似文本的指纹信息只有少量的不同, 因此可以通过计算文本的指纹信息即 Simhash 值的相似程度来判断文本的相似程度. 海明距离表示两个文本 Simhash 值每一个索引位置值不同的数量, 假设两个文本$(a_1,a_2,\cdots,a_n)$与$(b_1,b_2,\cdots,b_n)$的 Simhash 值长度为n,

i表示第i位, 则文本a和b之间的海明距离计算公式为:

$$H(a,b)=\sum_{i=1}^{n}a_i\otimes b_i \tag{7}$$

其中, $\otimes$表示异或运算.

Simhash 算法中, 首先将文本信息映射得到指纹信息, 再通过计算海明距离$H(a,b)$来判断相似度. 在实践中, 通常认为两个文本的海明距离$H(a,b)\leqslant 3$时文本是相似的, 本文采用$H(a,b)=3$作为判断相似性的阈值. 海明距离$H(a,b)$是文本评级模型的指标之一.

2.2 Shingling 算法

Shingling 算法是一种降低特征维度去检测文本相似性的方法[18]. Shingling 算法是将文本的相似性转化为词语集合的相似性, 首先将文本M划分成一些大小为w的连续子序列的集合$(w_1,w_2,\cdots,w_n)$称为$S(M,w)$, 再通过两个集合的交集除以并集的计算方式表示文本的相似性[19,20], 则文本A和B的相似性定义为:

$$r_w(A,B)=\left|\frac{S(A,w)\cap S(B,w)}{S(A,w)\cup S(B,w)}\right| \tag{8}$$

式 (8) 得到的$r_w(A,B)$是文本的相似系数即文本A和B的重复率, 重复率是标书文本评估的指标之一.

3 评估指标

基于文本分析的标书综合评估模型由文本评估模型和文本评级模型组成, 两者的计算结果共同实现了对标书的综合评估.

3.1 文本评估模型

在传统的评标中, 通常是评标专家对投标文件的 3 个主要部分: 商务标部分、技术标部分和报价部分进行打分, 每部分得分与权重相乘后累加即可得到专家评分结果. 在这个过程中, 围标、串标行为的识别往往依靠评分专家的主观判断, 缺少客观的评定指标.

基于文本分析的标书综合评估模型分为标书文本评估和文本评级. 文本评估模型是在传统的评分指标“商务标部分X_1”“技术标部分X_2”“价格得分X_3”的基础上加入了基于 Shingling 算法计算得到的标书文本的“重复率X_4”和投标文件要求的招标文件目录与真实目录的“匹配度X_5”. 其中X_1、X_2得分是参考专家经验计算得到的, X_3、X_4、X_5是模型评分.

本实验中标书的评标基准价采用平均值法, 评标基准价的计算方法[21] 为:

C(评标基准价)=A(所有有效标书的平均价格) (9)

价格得分X_3的计算方法是: 当投标报价=C时, 该标书的价格分为 100 分; 投标报价>C时, 每高于评标基准价 1%, 扣 1 分; 投标报价<C时, 每低于评标基准价 1%, 扣 0.5 分. 价格得分<0 时, 记 0 分.

重复率指标X_4的计算方法是: 当重复率$F\leqslant n$%时, 得 100 分; 当重复率$F>n$%时, 每高于重复率 1%, 扣 2 分, 本文取n=5. 当重复率得分<0 时, 记 0 分.

目录匹配度指标X_5是把招标文件对于投标文件的要求目录中的各级标题提取关键词, 利用关键词与真实目录进行字符串匹配, 匹配度的计算公式为:

$$P=\frac{n(\text{成功匹配数})}{n(\text{总数})}\times 100\% \tag{10}$$

匹配度X_5的计算方法是: 当P=100% 时, 得 100 分; 当P <100%, 每少 1%, 扣 5 分. 当匹配度得分<0 时, 记 0 分.

为了获取到指标X_i在评标过程中所占的权重, 本实验通过调查问卷的方法收集了 5 位评标专家对 5 个指标中两两相比时的相对重要性排序, 然后通过层次分析法 (analytic hierarchy process, AHP) 获取了指标X_i对应的权重a_i. 文本评估得分的计算公式为:

$$X=\sum_{i=1}^{5}X_i\times a_i \tag{11}$$

文本评估模型通过对 5 项指标进行定量计算, 得到了指标权重, 并进一步得到各标书的得分. 文本评估模型的指标权重是基于评标专家的经验, 采用半定量的层次分析法确定. 文本评估模型的各项指标综合了传统评估指标、重复率和目录匹配度, 是一种更加全面的评价方法, 具有一定的通用性. 文本评估模型的指标列表见表 1.

3.2 文本评级模型

在传统的评标中, 招投标行为是否出现围标、串标行为往往是通过评标专家现场进行人工识别, 一方面效率较低且难以发现围标、串标行为的有效证据, 另一方面, 人工难以有效的挖掘标书的深层次信息.

文本评级模型识别围标串标行为的方法是通过 7 个指标对标书分别进行评级, 进行风险等级划分, 最终综合 7 个评级结果, 通过综合评级式 (18) 得到最终的标书评级, 7 个指标分别是: 利用基于权重改进的 Simhash 算法得到的两个标书文本全文之间的相似度

指标Y_1和标书文本的技术标部分 (技术标是投标的关键性内容) 的相似度指标Y_2. 定义$H(A,B)$(即海明距离)为标书文本A和B之间的相似度, 则Y_1和Y_2评级公式为:

$$Y_1/Y_2=\begin{cases}1\text{ (正常)}, & H(A,B)>10\\ 0\text{ (疑似)}, & 3<H(A,B)<10\\ -1\text{ (高度疑似)}, & H(A,B)\leqslant 3\end{cases}\tag{12}$$

表 1 文本评估模型的指标列表

评估类型	评估指标	符号	评分	属性	权重	得分
文本评估	商务标得分	X_1	人工评分	分数: 依据评标细则	a_1 : 18%	百分制
	技术标得分	X_2	人工评分	分数: 依据评标细则	a_2 : 45%	
	价格得分	X_3	模型评分	分数: 依据公布的基准价	a_3 : 27%	
	重复率	X_4	模型评分	百分数: 基于Shingling算法	a_4 : 6%	
	目录匹配度	X_5	模型评分	百分数: 目录匹配数/目录总数	a_5 : 4%	

基于知识图谱的投标企业关联度指标Y_3, 通过外部系统中的知识图谱获取两个企业间工程建设项目、资金等信息往来情况. 定义$C(A,B)$为企业A和企业B的项目往来次数, 则Y_3的评级公式为:

$$Y_3=\begin{cases}1\text{ (正常)}, & C(A,B)\leqslant 2\\ 0\text{ (疑似)}, & 2<C(A,B)<5\\ -1\text{ (高度疑似)}, & C(A,B)>5\end{cases}\tag{13}$$

投标企业的企业资质与投标价格的一致性指标Y_4, 通过投标企业的标书获取报价、总资产、已完成同类项目数量, 定义投标企业的报价排序为a、企业总资产排序为b和已完成同类项目数量排序为c, 则Y_4评级公式为:

$$Y_4=\begin{cases}1\text{ (正常)}, & |a-b|+|a-c|\leqslant 3\\ 0\text{ (疑似)}, & 3<|a-b|+|a-c|\leqslant 5\\ -1\text{ (高度疑似)}, & |a-b|+|a-c|>5\end{cases}\tag{14}$$

投标价的价格上 (下) 浮率指标Y_5, 文献 [22] 验证了围标、串标的企业通常由一定数量的相同或相似报价的企业和一定数量的远低于正常报价的企业共同组成, 这些企业从价格方面使得评标基准价向组织围标、串标的企业靠近. 本实验的投标价A相对于基准价C价格上 (下) 浮率为$F(A,C)$, 则Y_5的评级公式为:

$$Y_5=\begin{cases}1\text{(正常)}, & 5\%<|F(A,C)|<20\%\\ 0\text{ (疑似)}, & 2\%<|F(A,C)|\leqslant 5\%\\ -1\text{ (高度疑似)}, & \text{其他}\end{cases}\tag{15}$$

基于政府信息公开的投标企业诚信度指标Y_6, 通过政府公开信息查询网站获取投标企业的社会信用情况、资金状况和违法违规情况的负面记录数量$J(A)$, 并进行评级, Y_6的评级公式为:

$$Y_6=\begin{cases}1\text{ (正常)}, & J(A)=0\\ 0\text{ (疑似)}, & 1\leqslant J(A)\leqslant 3\\ -1\text{ (高度疑似)}, & J(A)>3\end{cases}\tag{16}$$

基于异常行为的指标Y_7, 异常行为是指: 文件混装、未按照要求撰写投标文件等故意废标的情况; 不同标书的签名字迹一致; 截标前多家企业同时撤回标书; 不同企业的投标保证金出自同一账户等. 企业A异常行为的数量记为$M(A)$, 其Y_7评级公式为:

$$Y_7=\begin{cases}1\text{ (正常)}, & M(A)=0\\ 0\text{ (疑似)}, & 1\leqslant M(A)\leqslant 3\\ -1\text{ (高度疑似)}, & M(A)>3\end{cases}\tag{17}$$

对于评级指标$Y_1Y_2,\cdots,Y_7$, 每个评级指标根据评级公式可以得到对应的评级数, 将每个指标的评级构造评级向量$\vec{r}$, 其中Y_i的评级值是评级向量$\vec{r}$第i位的值, 例如$\vec{r}=(1,-1,0,1,0,1,1)$.

评级指标Y_1,Y_2,Y_3是衡量标书文本A与标书文本B之间的关系的评级, 其他指标是标书文本A自身的评级, 则定义标书文本评级公式:

$$Y_{r_i}=\begin{cases}\text{正常}, & R(r_i=1)\geqslant 5\\ \text{疑似}, & \text{其他}\\ \text{高度疑似}, & R(r_i=1)\leqslant 3\text{或}R(r_i=-1)\geqslant 3\end{cases}\tag{18}$$

其中, $R(\cdot)$表示所有可能的情况中使得括号内条件成立的情况的个数.

文本评估模型的创新之处在于该模型考虑了文本层面的分析、企业关联分析、背景分析与行为分析等因素, 通过 7 个指标的评级结果综合分析得到识别围标串标的结果, 为围标、串标行为的检测提供了支撑, 文本评级指标列表见表 2.

3.3 综合评估

标书评估实践中, 最重要的两个步骤是对标书进行评分得到排序和识别围标、串标行为, 从而确定最终入围的标书. 但在技术研究中, 往往只少数文献对某些方面进行了分析, 并未考虑到标书分析的全面性和客观性问题.

表 2 文本评级模型的指标列表

评估类型	评估指标	符号	评分	属性 (评级分别对应1; 0; –1)
文本评级 (正常; 疑似; 高度疑似)	全文相似度	Y_1	模型评分	定性: 低; 中; 高
	技术标部分相似度	Y_2	模型评分	定性: 低; 中; 高
	企业关联度	Y_3	人工评分	定性: 低; 中; 高
	资质与报价一致性	Y_4	模型评分	定性: 高; 中; 低
	价格上(下)浮率	Y_5	模型评分	定性: 低; 中; 高
	企业诚信度	Y_6	人工评分	定性: 高; 中; 低
	异常行为	Y_7	人工评分	定性: 低; 中; 高

本文提出的标书评估模型分别从文本评估 (指标 X_1–X_5) 和文本评级 (指标 Y_1–Y_7) 两个方面进行标书的定量计算和定性分析. 文本评估 (X项) 是在传统的评分指标中加入了“重复率X_4”和“目录匹配度X_5”, 并利用层次分析法获得指标对应的权重, 从定量计算方面实现对文本的评分, 确认投标企业的标书得分排序. 文本评级 (Y项) 是利用 7 项指标的定性评级结果判断投标企业是否出现疑似围标、串标的行为, 结合人工进行核查, 为文本评估 (X项) 提供围标、串标的参考, 两者共同实现对标书的综合评估.综合评估本质是将两个不同方面的计算结果进行结合, 但是这种结合又加入了人工的核查, 增大了模型的准确性和可靠性.

4 实际案例分析

本节将中国水利电力物资集团有限公司工程建设中两个招投标项目的文本和数据作为实际案例数据进行实验, 通过基于文本分析的标书综合评估模型的计算结果与真实结果进行对比, 展示本文模型的有效性.

在文本评估中的指标X_4 (重复率), 文本评级中的指标Y_1 (全文相似度)、Y_2 (技术标部分相似度)、Y_3 (企业关联度) 是描述两文本之间的关系, 当某项目有$A_1, A_2, \cdots, A_n$共n个企业进行投标, 在计算A_i的这 4 个指标时, 要将A_i与其他$n-1$个企业进行比较, 共有$n-1$个结果, 结果应当选择数值属性最不利于该企业的实验数据作为A_i在该指标的数据值. 此外, 实验数据中的招标项目的投标企业个数通常为 4–8 个, 所以计算的复杂度是合理的.

实际案例数据分别采用“某电厂入厂次干道”项目和“某电站公用及辅机控制设备”项目的案例数据. “某电厂入厂次干道”招标项目共有 4 家企业进行投标, 即共有 4 份标书文本. 经过标书文本的数据处理得到“某电厂入厂次干道”项目的指标数据与综合评估结果, 见表 3. 其中有 3 家企业的标书被识别为“正常”, 1 家企业的标书被识别为“疑似围标、串标”. 投标企业 4 被识别为“疑似围标、串标”, 这是由于投标企业 4 的文本评级结果中有两项评级为“–1”, 根据文本评级式 (18), 故被识别为“疑似围标、串标”.

“某电厂入厂次干道”项目的招标文件规定了根据评标分数选择评分最高的 3 家企业作为“晋级”企业. 实验数据也采用评分排序前 3 的企业为“晋级”企业, 进入候选标书名单. 经过综合评估结果与专家评标结果和评标报告进行对比, 发现实验评分结果与专家评标的真实评分结果吻合, 识别围标、串标结果为评分结果提供参考, 为人为识别围标、串标行为提供依据, 实验结果见表 4.

对“某电站公用及辅机控制设备”项目标书文本进行处理, 项目共有 6 家企业进行投标, 经过标书文本的处理, 最终得到“某电站公用及辅机控制设备”项目的指标数据与综合评估结果分析, 见表 5. 经过综合评估结果与专家评标结果和评标报告进行对比, 发现实验结果与专家的真实结果吻合, 实验结果见表 6.

通过 2 个项目共 10 个标书的案例分析, 并将实验结果与真实结果进行对比, 发现通过标书综合评估模型的计算结果与真实结果吻合, 表明了基于文本分析的标书综合评估模型的在本节 2 个项目案例分析上的有效性. 该模型的评估从定量计算和定性分析两个方面分别实现了文本评估和文本评级, 两者的结果共同构成了综合评估的结果. 在实践中, 文本评估模型为专家打分提供数据支持, 提高了人工评标的效率; 文本评级模型能够为招投标过程中围标、串标行为的识别提供依据, 大大提升识别围标、串标行为的效率和效果, 识别为疑似或高度疑似存在围标、串标行为的企业标书需进行人工核查, 得到围标、串标行为的识别结果.

表 3 “某电厂入厂次干道”项目的指标数据与综合评估结果

评估模型	评估指标	符号	权重/评级	得分/评级			
				投标企业1	投标企业2	投标企业3	投标企业4
文本评估	商务标得分	X_1	18%	93.33	95.00	92.33	83.67
	技术标得分	X_2	45%	90.67	87.33	93.00	90.33
	价格得分	X_3	27%	98.81	99.68	87.75	76.06
	重复率	X_4	6%	88.40	93.80	99.00	88.40
	目录匹配度	X_5	4%	100.00	100.00	100.00	100.00
文本评级	全文相似度	Y_1	1;0;−1	1	1	1	1
	技术标相似度	Y_2	1;0;−1	1	0	1	0
	企业关联度	Y_3	1;0;−1	1	1	1	1
	资质与报价一致性	Y_4	1;0;−1	1	1	0	−1
	价格上(下)浮率	Y_5	1;0;−1	−1	−1	−1	−1
	企业诚信度	Y_6	1;0;−1	1	1	1	1
	异常行为	Y_7	1;0;−1	1	1	1	1
综合评估	—			93.583 6 正常	92.940 1 正常	92.101 9 正常	85.549 3 疑似

表 4 “某电厂入厂次干道”项目模型数据和真实数据表

企业	实验数据	实验结果	真实数据	真实结果	结果对比
企业1	93.58; 正常	晋级	92.00; 正常	晋级	吻合
企业2	92.94; 正常	晋级	89.67; 正常	晋级	吻合
企业3	92.10; 正常	晋级	93.67; 正常	晋级	吻合
企业4	85.55; 疑似	淘汰	85.60; 正常	淘汰	吻合

表 5 “某电站公用及辅机控制设备”项目的指标数据与综合评估结果

评估模型	评估指标	符号	权重/评级	得分/评级					
				投标企业1	投标企业2	投标企业3	投标企业4	投标企业5	投标企业6
文本评估	商务标得分	X_1	18%	90.34	90.00	88.67	82.67	83.00	91.34
	技术标得分	X_2	45%	86.34	86.33	90.33	86.33	87.33	92.00
	价格得分	X_3	27%	96.40	99.10	91.10	90.70	99.00	90.10
	重复率	X_4	6%	100.00	100.00	100.00	100.00	100.00	100.00
	目录匹配度	X_5	4%	100.00	100.00	100.00	100.00	100.00	100.00
文本评级	全文相似度	Y_1	1;0;−1	1	1	1	1	1	1
	技术标相似度	Y_2	1;0;−1	1	0	1	1	1	1
	企业关联度	Y_3	1;0;−1	1	1	1	1	1	1
	资质与报价一致性	Y_4	1;0;−1	1	0	1	−1	1	−1
	价格上(下)浮率	Y_5	1;0;−1	0	−1	1	1	−1	1
	企业诚信度	Y_6	1;0;−1	1	1	1	1	1	1
	异常行为	Y_7	1;0;−1	1	1	1	1	1	1
综合评估	—			91.14 正常	91.81 疑似	91.21 正常	88.22 正常	90.97 正常	92.17 正常

表 6 “某电站公用及辅机控制设备”项目模型数据和真实数据表

企业	实验数据	实验结果	真实数据	真实结果	结果对比
企业1	91.14; 正常	淘汰	89.00; 正常	淘汰	吻合
企业2	91.81; 疑似	晋级	88.34; 正常	晋级	吻合
企业3	91.21; 正常	晋级	89.67; 正常	晋级	吻合
企业4	88.22; 正常	淘汰	84.67; 正常	淘汰	吻合
企业5	90.97; 正常	淘汰	86.34; 正常	淘汰	吻合
企业6	92.17; 正常	晋级	91.67; 正常	晋级	吻合

5 结论与展望

招投标是工程建设中的重要环节, 高效地识别围标、串标行为是招投标过程的一大难题, 在实践领域人工识别围标、串标行为效率较低、成本高, 在研究领域缺少全面、完善的评估方法. 本文的创新点在于提出了融合文本评估和文本评级的综合评估模型, 模型基于定量计算和定性分析两个方面进行标书处理, 同时将 Shingling 算法和改进的 Simhash 算法用于标书文本分析之中. 通过建立基于文本分析的标书综合评估模型, 提取文本的数据信息, 对标书建立文本评估模型和文本评级模型, 实现了对标书的定量和定性的分析, 进而实现对标书的综合评估. 该模型不仅能够为标书评估提供更加客观、合理的得分依据, 为识别投标企业围标串标行为提供有效的参考, 还能提高标书评分的效率. 除此之外, 也能为构建电子化招投标系统和建立标书分析模型提供条件与准备. 基于文本分析的标书综合评估模型对工程建设项目中的标书评标工作具有重要的意义, 基于标书数据形成的知识图谱也为电子化招投标中属性关系的建立和未来的深度探索提供有力的支撑.

基于文本分析的标书综合评估模型仍可在以下几个方面进行改进: 首先, 随时招投标领域向电子化方向发展, 标书评估中用到的评估指标还需要根据国家政策法规、招投标实际情况、招投标工程领域等方面进行补充和完善; 其次, 针对较多数量的标书, 需要采取更加高效、快速的方法识别文本之间的相似度和重复率; 最后, 需要采取不同的方法论证本文模型的有效性和可解释性.

参考文献

1 聂军民. 企业招标投标工作中存在的问题及对策. 现代商贸工业, 2021, 42(5): 118–119. [doi: 10.19311/j.cnki.1672-3198.2021.05.058]

2 张坤, 唐勇, 胡剑炜. 电子化招投标中围标串标防治措施. 招标采购管理, 2020, (3): 42–43. [doi: 10.3969/j.issn.2095-4123.2020.03.015]

3 白臻. 工程项目招投标围标串标防范对策研究. 中国科技投资, 2020, (5): 179–180.

4 王志强, 邵良杉. 基于 AHP 的标书模糊综合评价方法. 科技情报开发与经济, 2007, 17(9): 176–178. [doi: 10.3969/j.issn.1005-6033.2007.09.105]

5 奚雪峰, 周国栋. 面向自然语言处理的深度学习研究. 自动化学报, 2016, 42(10): 1445–1465. [doi: 10.16383/j.aas.2016.c150682]

6 韩程程, 李磊, 刘婷婷, 等. 语义文本相似度计算方法. 华东师范大学学报 (自然科学版), 2020, (5): 95–112. [doi: 10.3969/j.issn.1000-5641.202091011]

7 邵恒, 冯兴乐, 包芬. 基于深度学习的文本相似度计算. 郑州大学学报 (理学版), 2020, 52(1): 66–71, 78. [doi: 10.13705/j.issn.1671-6841.2019007]

8 Gharghe ZE, Bidgoli BM. Weighted shingling: An adaptation of shingling for weighted shingles. Proceedings of 2009 IEEE International Conference on Innovations in Information Technology (IIT). Al Ain: IEEE, 2009. 150–154.

9 Ye SZ, Wen JR, Ma WY. A systematic study on parameter correlations in large-scale duplicate document detection. Knowledge and Information Systems, 2008, 14(2): 217–232. [doi: 10.1007/s10115-007-0071-9]

10 陈二静, 姜恩波. 文本相似度计算方法研究综述. 数据分析与知识发现, 2017, 1(6): 1–11.

11 马成前, 毛许光. 网页查重算法 Shingling 和 Simhash 研究. 计算机与数字工程, 2009, 37(1): 15–17, 108. [doi: 10.3969/j.issn.1672-9722.2009.01.004]

12 张庆颖. 基于 Simhash 和 CNN 的相似新闻推荐 [硕士学位论文]. 成都: 电子科技大学, 2020.

13 余意, 张玉柱, 胡自健. 基于 Simhash 算法的大规模文档去重技术研究. 信息通信, 2015, (2): 28–29.

14 董博, 郑庆华, 宋凯磊, 等. 基于多 SimHash 指纹的近似文本检测. 小型微型计算机系统, 2011, 32(11): 2152–2157.

15 Manku GS, Jain A, Sarma AD. Detecting near-duplicates for web crawling. Proceedings of the 16th International Conference on World Wide Web. Banff, Alberta: ACM, 2007. 141–150.

16 张航, 盛志伟, 张仕斌, 等. Simhash 算法在文本去重中的应用. 计算机工程与应用, 2020, 56(11): 246–251. [doi: 10.3778/j.issn.1002-8331.1902-0246]

17 Shivakumar N, Garcia-Molina H. Finding near-replicas of documents on the Web. Proceedings of International Workshop WebDB'98. Valencia: Springer, 1998. 204–212.

18 于建坤. 云环境下搜索引擎系统关键技术研究 [硕士学位论文]. 南京: 南京邮电大学, 2016.

19 郝忠翁. 大规模 Web 文本快速分类关键技术研究 [硕士学位论文]. 哈尔滨: 哈尔滨工程大学, 2015.

20 葛慧. 相似性的块级重复数据删除算法的研究 [硕士学位论文]. 沈阳: 辽宁大学, 2018.

21 徐原, 王蕾. 优化价格评分方法防范围标串标风险. 招标采购管理, 2020, (3): 44–47. [doi: 10.3969/j.issn.2095-4123.2020.03.016]

22 王莹. 建设工程中价格偏离淘汰法与围标串标的关系研究 [硕士学位论文]. 深圳: 深圳大学, 2018.