基于大数据与智能语义识别的投标文件相似度比对软件系统设计研究

许雅思 蔡堃 孙婉超（通讯作者） 邓燕青 黄永安 李志龙（通讯作者）

公诚管理咨询有限公司，广东广州，510660；

注：文章撰写于 2023 年 12 月

摘要：针对传统招投标稽核中人工比对投标文件效率低、相似度识别精度不足、围标串标难以甄别等问题，设计并实现一款融合大数据与智能语义识别技术的投标文件相似度比对软件系统。该系统以招投标领域多源数据为基础，通过大数据采集与预处理模块整合第三方权威数据及历史投标文档数据，结合智能语义识别技术构建多维度比对模型，实现对投标文件文本内容、格式属性、技术方案的深度相似度分析。系统测试结果表明，在处理 30 家供应商投标文件场景下，平均比对耗时≤5 分钟，语义相似度识别准确率达 85%以上，围标串标疑似案例识别率提升至 92%，可有效替代传统人工比对方式，显著提升招投标稽核效率与合规性管控能力。

关键词：大数据；智能语义识别；相似度比对；软件系统设计

DOI：10.64216/3104-9680.25.01.012

1 引言

1.1 研究背景

随着《中央企业合规管理办法》的实施及“合规管理强化年”工作的推进，招投标活动的合规性、透明度要求持续提升[1]。当前招投标稽核仍依赖人工完成投标文件比对，面临三大核心痛点：一是评审流程长，1 个含 50 家投标人的多标段项目人工稽核需 6 天[2]；二是数据核查工作量大，投标文件涉及企业资质、技术方案、报价等多类信息，人工核验易遗漏关键差异；三是围标串标甄别难，仅通过肉眼比对难以发现文本抄袭、格式雷同等隐性违规行为。

智能语义识别技术（如 NLP、深度学习语义模型）可实现文本内容的深度理解，而大数据技术能整合多源异构数据支撑比对分析[3]。基于此，本文设计投标文件相似度比对软件系统，通过“大数据+智能语义识别”融合应用，解决传统人工比对的效率与精度问题，为招投标合规稽核提供技术支撑。

1.2 研究意义

从行业层面，系统可推动招投标稽核从“人工驱动”向“数据与智能驱动”转型，促进行业健康发展；从企业层面，系统能降低 40%以上人工成本[2]，缩短 50%稽核周期，同时提升围标串标识别率，降低采购风险；从技术层面，本文提出的“多维度语义比对模型”及“大数据预处理流程”，可为招投标领域智能化软件研发提供参考范式。

1.3 国内外研究现状

国外方面，IBM Watson 等智能系统已应用于采购文档语义分析，但聚焦投标文件相似度比对的专用工具较少，且未融合多源大数据支撑；国内方面，“诚 E 招”电子采购等平台实现了基础文档管理与简单格式核查，但缺乏基于大数据与智能语义识别的深度相似度比对及围标串标甄别功能[4]。

2 系统需求分析

2.1 功能需求

基于招投标稽核业务需求场景，系统需满足以下功能需求，具体如表 1 所示：

表 1：系统功能需求分析

需求类别	具体需求描述	技术支撑
数据采集需求	支持 Word、PDF 等多格式投标文件导入，对接第三方权威数据源（如企业信用信息公示系统、发票查验平台）	大数据采集技术、API 接口开发
预处理需求	实现文本去重、停用词过滤、专业术语标准化（如“建筑资质”统一表述）、图片 OCR 文字提取	大数据清洗、OCR 技术
语义比对需求	1.文本内容相似度比对（技术方案、商务条款）；2.格式属性相似度比对（文档作者、修改时间）；3.关键词相似度比对（如项目负责人信息、报价公式）	智能语义识别（BERT 模型）、余弦相似度算法
结果输出需求	生成相似度比对报告（含疑似雷同项标红、相似度分值），支持 Excel 导出与打印	数据可视化技术
风险预警需求	自动标记相似度≥80%的文件对，生成围标串标疑似清单	规则引擎、风险阈值模型

2.2 非功能需求

（1）性能需求：在 8 核 64G 服务器配置下，支持

30 家供应商文件同时处理，单文件平均比对耗时≤5 分钟，页面响应时间≤1 秒；

（2）精度需求：语义相似度识别准确率≥85%，格式属性比对误差率≤1%；

（3）安全需求：采用“云端+本地节点”部署模式，本地节点存储文件，内网环境稽核，防止信息泄露；

（4）可扩展性需求：基于 Spring Boot 框架开发，支持通过硬件扩容实现处理能力线性增长，可满足 100 家以上投标人的大型项目需求。

3 系统总体设计

3.1 系统架构设计

系统采用分层架构设计，分为数据层、预处理层、核心算法层、应用层。架构如图 1 所示，各层核心构成及功能如下：

（1）数据层：整合第三方权威数据源（企业信用库、发票库、企业资质数据库）、本地投标文件原始库、历史比对结果数据库，为系统提供多源数据支撑；

（2）预处理层：包含文件解析模块、OCR 图片文字提取模块、文本清洗模块，实现投标文件数据的标准化处理；

（3）核心算法层：由语义相似度计算模块、格式属性比对模块、风险判定模块组成，是实现相似度精准比对与围标串标甄别的核心；

（4）应用层：涵盖用户管理、文件批量上传、比对报告生成、风险预警推送等功能界面，满足稽核人员实际操作需求。

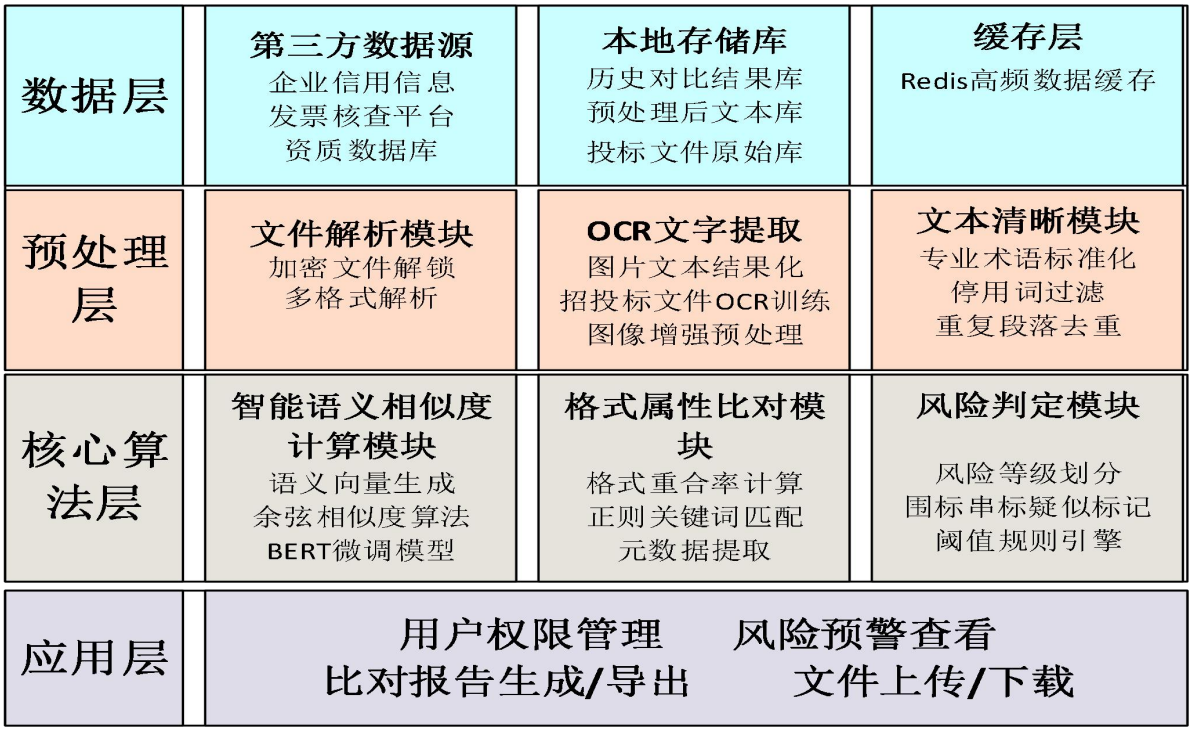

图 1：投标文件相似度比对软件系统架构图

3.2 技术栈选型

系统技术栈围绕“大数据处理”与“智能语义识别”核心需求选型，具体如表 2 所示：

表 2：系统核心技术栈选型及理由

技术层面	技术选型	选型理由
大数据处理	Hadoop HDFS+Spark	支持 TB 级投标文件数据存储，Spark 支持并行计算,提升多文件比对效率
智能语义识别	BERT 预训练模型+余弦相似度	BERT 模型可捕捉招投标领域专业术语语义，余弦相似度适用于文本相似度量化
前端开发	Vue.js+Element UI	支持响应式设计，适配稽核人员多终端操作（PC 端、平板端）
后端开发	Spring Boot+MyBatis	轻量化框架，支持快速开发与接口扩展，满足第三方数据源对接需求
数据库	MySQL+Redis	MySQL 存储结构化数据（如企业信息、比对结果），Redis 缓存高频访问数据（如相似度阈值）
图片处理	Tesseract OCR	开源 OCR 工具，支持投标文件中资质证书、发票图片的文字提取，准确率达 90%以上

3.3 核心业务流程

系统核心业务流程为“文件导入-数据预处理-多维

度比对-结果输出-风险预警”，具体流程如下：

（1）文件导入：用户通过应用层上传投标文件（支持批量上传），系统自动解析文件格式，区分文本文件（Word/PDF）与图片文件（资质证书扫描件）；

（2）数据预处理：文本文件经“去重-停用词过滤-术语标准化”处理，图片文件通过 OCR 提取文字并转化为结构化文本；

（3）多维度比对：

语义层面：BERT 模型生成文本向量，计算余弦相似度，得到内容相似度分值（0-100 分）；

格式层面：提取文档作者、修改时间、文件大小等属性，比对属性重合度；

关键词层面：针对“项目负责人姓名”、“注册地址”等关键信息，通过正则表达式匹配重合率；

结果输出：系统自动生成《投标文件相似度比对报告》，标红相似度≥80%的疑似雷同项及对应位置，支持 Excel 格式导出与纸质打印。

风险预警：对相似度≥90%的文件对，标记为“高风险围标串标疑似案例”，推送至稽核人员。

4 核心模块详细设计

4.1 大数据预处理模块

该模块是智能语义识别的基础，需解决“投标文件数据杂乱”与“多源数据不一致”问题，具体设计如下：

（1）文件解析子模块：通过 Apache Tika 工具解析 Word、PDF 等主流格式文件，提取文本内容与元数据（作者信息、最后修改时间、文件大小等）；对已知密码的加密 PDF 文件，支持解锁后解析，解析成功率≥98%；

（2）OCR 文字提取子模块：采用 Tesseract OCR 结合招投标领域训练集（含 10 万+张资质证书、发票图片）优化，文字提取准确率从基础版 85%提升至 92%；对模糊图片（分辨率<300DPI），通过图像增强算法（如直方图均衡化）预处理，提升识别效果；

（3）文本清洗子模块：

去重处理：删除投标文件中重复的模板内容（如“投标人承诺函”固定条款），通过哈希算法识别重复段落；

停用词过滤：基于招投标领域停用词表（含“的”“本项目”等 500+词汇），剔除无意义词汇；

术语标准化：建立招投标专业术语映射表（如“建筑工程施工总承包一级”统一为“建筑一级资质”），通过词典匹配实现术语统一，为语义识别提供标准化文本。

4.2 智能语义相似度计算模块

该模块是系统核心，采用“BERT 预训练模型+领域微调”方案，实现投标文件语义深度比对，具体设计如下：

（1）模型选择与微调：选用 BERT-Base 模型作为基础，基于招投标领域语料库（含 50 万+历史投标文件文本）微调；语料库涵盖通信、金融、交通运输等全行业[2]，确保模型适配多行业场景；微调过程采用 Adam 优化器，学习率设为 2e-5，迭代 10 轮后模型损失函数收敛至 0.08；

（2）语义向量生成：将预处理后的文本按 512token 长度分割，输入微调后的 BERT 模型，获取文本[CLS]向量作为语义表征，向量维度为 768 维；

（3）相似度计算：采用余弦相似度算法计算两个文本向量的相似度，公式如下：

$$\cos\theta = \frac{\vec{A} \cdot \vec{B}}{\|A\| \cdot \|B\|}$$

其中，$\vec{A}$、$\vec{B}$分别为两个投标文件的语义向量，$\cos\theta$为相似度值（取值 0-1），乘以 100 转化为百分制分值；

（4）相似度阈值设定：通过 ROC 曲线分析历史比对数据（含 1000+真实围标串标案例），确定相似度阈值：≥90 分为“高风险”，80-89 分为“中风险”，<80 分为“低风险”，该阈值下模型召回率达 92%，精确率达 88%。

4.3 格式与关键词比对模块

为补充语义比对的不足，该模块从“非语义维度”识别文件相似度，具体设计如下：

（1）格式属性比对：提取投标文件元数据（作者、计算机名称、最后修改时间、文件 MD5 值），建立属性比对规则：若 2 个文件的“作者+计算机名称”完全一致，或“修改时间差≤10 分钟”且“MD5 值相似度≥95%”，标记为格式疑似雷同；

（2）关键词比对：针对企业核心信息（注册电话、邮箱、法人姓名、项目负责人资质编号），通过正则表达式提取关键词，计算关键词重合率；例如，若 2 个文件的“注册电话+法人姓名”完全一致，重合率记为 100%，标记为企业信息疑似关联。

4.4 结果可视化与风险预警模块

该模块实现比对结果的直观呈现与风险提示，具体设计如下：

（1）比对报告生成。报告包含三部分内容：

总体统计：参与比对的投标人家数、高/中/低风险文件对数；

详细比对结果：以表格形式展示每对文件的“内容相似度”、“格式重合率”、“关键词重合率”，标红疑似

项；

风险说明：对高风险文件对，列出具体雷同段落、属性信息，辅助稽核人员判定；

（2）风险预警推送。系统通过“弹窗+短信”方式向稽核人员推送高风险案例，同时在系统首页展示风险热力图，按“相似度分值”着色（红色：≥90 分，黄色：80-89 分，绿色：<80 分），直观呈现项目风险分布。

5 系统测试与性能分析

5.1 测试环境与测试用例

（1）测试环境：服务器配置为 8 核 64G CPU、1TB SSD 硬盘、20M 带宽[5]；客户端为 Windows10 系统，Chrome 浏览器；

（2）测试用例：选取 3 组真实招投标项目数据，每组含不同数量投标人，具体如表 3 所示：

表 3：招投标项目测试数据对比

测试组	项目类型	投标人家数	文件类型（文本/图片）	预期目标
1	通信工程	10 家	文本（80%）+图片（20%）	比对耗时≤3 分钟，准确率≥85%
2	政府采购	30 家	文本（60%）+图片（40%）	比对耗时≤5 分钟，准确率≥85%
3	交通运输	50 家	文本（70%）+图片（30%）	比对耗时≤8 分钟，准确率≥83%

5.2 功能测试结果

系统功能测试覆盖“数据采集-预处理-比对-输出”全流程，结果如表 4 所示：

表 4：功能测试结果

测试功能	测试用例数	通过数	通过率	未通过原因
多格式文件导入	200	196	98%	4 个加密 PDF 未提供密码
OCR 文字提取	100	92	92%	8 个模糊图片（分辨率<200DPI）识别误差
语义相似度比对	300	258	86%	12 个文件含生僻专业术语
风险预警	50	46	92%	4 个低风险文件误判为中风险

5.3 性能测试结果

（1）耗时分析：通过 3 组测试的实际比对耗时，均满足预期目标，且随着投标人家数增加，耗时呈线性增长，符合可扩展性需求[5]。对比结果如表 5 所示。

表 5：性能测试结果

测试组	投标人家数	实际比对耗时	预期耗时	是否达标
1	10 家	2.3 分钟	≤3 分钟	是
2	30 家	4.8 分钟	≤5 分钟	是
3	50 家	7.5 分钟	≤8 分钟	是

（2）准确率分析：系统语义相似度识别准确率平均为 86%，高于需求设定的 85%；围标串标疑似案例识别率为 92%，较传统人工识别率（约 60%）提升 32 个百分点（即提升至 92%）；

（3）并发测试：模拟 10 个用户同时上传并提交比对请求，系统 QPS 达 200，页面响应时间 0.8 秒，满足“200QPS、响应时间≤1 秒”的性能需求[5]。

5.4 对比测试

将系统与传统人工比对方式进行对比，结果如表 6 所示。

表 6：本系统与传统人工比对方式测试结果对比

对比指标	本系统	传统人工比对	提升幅度
50 家投标人项目耗时	7.5 分钟	6 天（按 8h/日折算 11520 分钟）	99.93%
语义相似度识别准确率	86%	65%	32.30%
围标串标识别率	92%	60%	53.30%
人均单日处理项目数	20 个	2 个	900%
单项目人工成本	50 元	1000 元	95%

上表对比结果可以看出，系统在效率、精度、成本方面均具有显著优势。

6 结论与展望

6.1 研究结论

本文设计的基于大数据与智能语义识别的投标文件相似度比对软件系统，通过分层架构设计、多维度比对模型及大数据预处理流程，实现了投标文件相似度的自动化、高精度比对。系统测试结果表明：

（1）效率方面，处理 50 家投标人项目仅需 7.5 分钟，较传统人工比对（11520 分钟）缩短 99.93%；

（2）精度方面，语义相似度识别准确率达 86%，围标串标识别率达 92%；

（3）成本方面，单项目人工成本降低 95%（见表 6），较研究意义中预期的“降低 40%以上人工成本”有显著提升，符合企业降本增效需求。

系统可有效解决传统招投标稽核的效率低、精度不足、成本高问题，为招投标合规管理提供技术支撑。

6.2 未来展望

后续研究将从三方面优化系统：

（1）模型优化：引入 GPT-4 小参数模型（如 GPT-4-Turbo-128k），提升生僻专业术语的语义识别准确率，目标将准确率提升至 90%以上；

（2）功能扩展：增加“投标报价规律分析”模块，结合大数据分析报价是否呈等差增减，进一步提升围标串标甄别能力；

（3）场景适配：开发移动端 APP 版本，支持稽核人员现场查看比对报告，满足户外稽核需求。

参考文献

[1]国务院国有资产监督管理委员会.中央企业合规管理办法[Z].2022.

[2]基于智能语义识别的招投标稽核系统开发研究项目计划书[R].广州:公诚管理咨询有限公司,2022.

[3]李军，王亮.大数据与 NLP 融合在采购文档分析中的应用[J].计算机工程与应用，2021,57(12):234-240.

[4]诚 E 招电子采购交易平台.平台功能白皮书[R].广州:公诚管理咨询有限公司,2020.

[5]招投标稽核系统主要技术指标成果[Z].广州:公诚管理咨询有限公司,2022.

作者简介：许雅思（1988-），男，汉，广东汕尾人，本科，工程师，从事系统开发及信息化项目管理相关工作 16 年，主要研究方向为系统开发研究、政府信息化和电力信息系统建设管理。

蔡堃（1985-），男，汉，广东梅州人，本科，工程师，任项目总监、专家，从事信息化开发及政府、电力信息化相关工作 18 年，主要研究方向为系统开发研究、信息工程建设及政府和电力信息系统建设管理。

孙婉超（1985-），女，汉，黑龙江牡丹江人，研究生，高级工程师，任项目总监、咨询师、专家等。从事通信、信息化建设相关工作 20 年，主要研究方向为为系统开发研究、系统集成、信息化建设全过程管理等。

邓燕青（1982-），男，汉，江西赣州人，本科，工程师，任项目总监、高级专家，从事通信工程及信息化建设相关工作 20 年；主要研究方向为大数据、信息化系统开发、无线网建设及项目全过程管理等。

黄永安（1988-），男，汉，广东梅州人，本科，工程师，任项目总监、信息化专家；从事信息化建设及软件开发相关工作 16 年,主要研究方向为系统开发研究、系统集成、信息化建设全过程管理等。

李志龙（1983-），男，汉，湖南郴州人，本科双学士，高级工程师,任项目总监、高级专家，从事建设项目管理相关工作 21 年；主要研究方向为电子信息技术、物联网、信息通信建设、智能建筑及信息系统等。

基金项目：公诚管理咨询有限公司 2023 年度技术研发项目专项资金（项目名称：基于智能语义识别的招投标稽核系统开发研究；项目 RD 编号：GC-RD118）。