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） 刘萍 叶明川 李志龙（通讯作者）

公诚管理咨询有限公司，广东广州，510660；

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：10.69979/3041-0673.26.03.109

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。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；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。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

从行业发展角度来看，“智能语义识别技术”的实施，有助于重构招投标监督的运行逻辑。在传统模式下，监督效能受限于人力资源与主观经验，而引入大数据与智能语义识别后，决策依据逐步由定性判断转向定量分析。对于企业而言，这种转变直接体现在成本与效率两个维度：初步测算显示，人工投入可减少 40%以上，核查周期压缩约一半，同时对围标串标行为的识别能力明显增强，从而降低采购环节的合规风险。

此外，本文构建的多维度语义比对模型及相应的数据标准化流程，也为招投标领域的智能化工具研发提供了可借鉴的技术路径，具有一定的外溢价值。

1.3 国内外研究现状

在智能语义解析方面，以 IBM Watson 为代表的系统已在部分国家的采购文档分析中投入使用，显示出较强的文本理解与信息抽取能力。不过，现有工具大多聚焦于通用文档处理，针对投标文件相似度检测的场景化应用相对有限，尤其是未能充分整合企业信用、税务、资质等多源数据进行综合研判。反观国内，多数电子采购平台仍以流程管理为核心，功能集中在文档存储、格式校验等基础环节，尚未形成基于大数据与深度语义识别的系统性相似度检测机制，围标串标识别仍主要依赖人工经验。

2 系统需求分析

2.1 功能需求

在功能设计阶段，优先考虑的是系统能否嵌入现有稽核流程，而非单纯技术指标的提升。数据采集方面，除支持 Word、PDF 等主流格式外，还需具备与权威数据源对接的能力，如企业信用信息公示系统与发票查验平台，以保证基础数据的真实性与时效性。预处理环节则承担大量基础性工作，包括文本去重、停用词过滤、专

业术语归一化，以及对资质证书等扫描件的 OCR 识别。在此基础上，系统从文本内容、文件属性、关键字段三个维度展开比对分析。最终输出形式以可视化检测报告为主，对高相似片段进行标注并量化评分，同时预留 Excel 导出与纸质打印接口。风险预警模块则依据预设阈值（如相似度≥80%）自动生成疑似围标串标清单，供稽核人员进一步核实。

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）在性能层面，系统需在 8 核 64GB 内存的服务器配置下稳定运行，能够并行处理约 30 家供应商的投标文件。实际测试中，单项目的全量比对耗时控制在 5 分钟以内，前端界面响应时间基本维持在 1 秒以内，能够满足一线稽核工作的实时性要求。

2）精度方面，系统更关注“可用”而非“完美”。文本语义相似度的识别准确率设定在不低于 85%的区间，而对于文件格式属性的对比任务，误差率则需控制在 1%以内，以避免因数据漂移导致误判。

3）安全设计采取了混合部署策略，将云端的计算能力与本地节点的存储能力结合起来。所有投标文件均保留在内网环境中，核查过程不对外传输敏感数据，从而降低信息泄露风险。

4）扩展性则是系统设计时的另一项隐性约束。基于 Spring Boot 框架开发，使系统具备良好的横向扩展能力。通过增加硬件资源，处理能力可实现近似线性提升，从而支撑涉及百余家投标单位的大型复杂项目，而不必对核心代码进行大规模重构。

3 系统总体设计

3.1 系统架构设计

系统采用分层架构模式，划分为数据层、预处理层、核心算法层、应用层四个层级。其架构如图 1 所示：

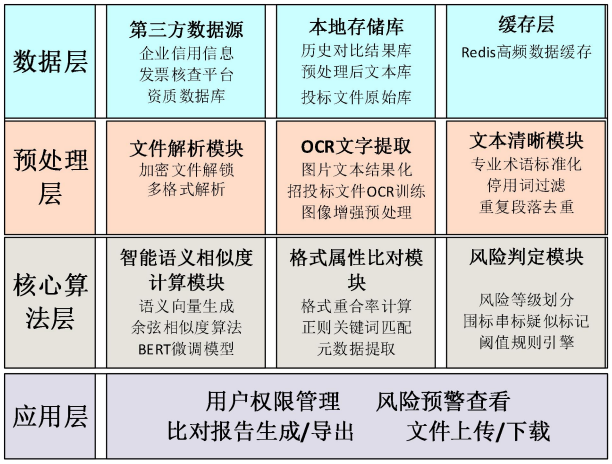

图 1：投标文件相似度比对软件系统架构图

1）数据层：其作用并不局限于存储，而是强调“通”与“融”。一方面，通过标准化接口对接企业信用、税务发票、资质信息等第三方数据库，保证数据来源的权威性；另一方面，将本地存储的投标文件原始数据、历史检测结果等内部资源纳入统一管理，为上层分析提供稳定的数据底座。

2）预处理层：该层更偏向“加工车间”的定位，承担文件解析、图像文字提取、文本清洗等任务。其目标不仅是格式转换，更是将形态各异的投标文件转化为可被算法直接使用的规范化数据。

3）核心算法层：真正决定系统智能水平的是核心算法层。该层由语义相似性计算、格式属性对比与风险判定三大模块组成，每一部分都针对不同类型的围标串标特征进行建模。正是这一层的存在，使系统能够从单

纯的“文件比对”升级为“行为识别”。

4）应用层：该层面向实际使用场景，提供用户管理、批量文件上传、检测报告生成及风险信息推送等交互功能。界面设计更多考虑稽核人员的操作习惯，避免过多技术细节暴露在前端，从而降低使用门槛。

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

整个系统的运行逻辑并非单一线性流程，而是围绕“从文件到风险信号”这一目标展开的递进式处理链条。用户首先在应用层完成投标文件的批量上传，系统会根据文件特征自动判别类型，将 Word、PDF 等可直接解析的文本文件，与资质证书、发票扫描件等图像类文件区分开来，以便进入不同的处理通道。

在数据准备阶段，文本文件会经历去重、无效词过滤以及术语归一化等清洗操作，而图像类文件则依赖 OCR 技术提取文字内容，并进一步转化为结构化文本。这一过程虽然看似前置，但实际上决定了后续所有比对结果的稳定性。

真正的核心计算集中在多维度对比环节。语义层面，系统借助 BERT 模型将文本映射为高维向量，并通过余弦相似度衡量文件间的语义接近程度；格式层面，则提取作者、修改时间、文件大小等元数据，用于捕捉潜在的批量生成痕迹；关键词层面引入正则表达式，对项目负责人、注册地址等关键字段进行精准抽取与重合度统计。三类结果并非孤立存在，而是在综合判定中共同发挥作用。

最终，系统会自动生成相似性检测报告，对相似度达到或超过 80%的文件组合进行明确标注，并提供 Excel 导出与纸质打印接口，以满足不同场景下的使用需求。在此基础上，风险预警机制会将相似度≥90%的案例标记为高风险对象，并直接推送至核查人员，以便在人工复核阶段优先处置。

4 核心模块详细设计

4.1 大数据预处理模块

在智能语义识别体系中，预处理模块承担着“打地基”的角色，其核心任务是缓解投标文件来源多样、数据结构不一致所带来的干扰。文件解析环节并未局限于单一技术手段，而是基于 ApacheTika 构建通用解析框架，能够稳定处理 Word、PDF 等常见格式，并同步提取作者、修改时间、文件大小等元数据。对于设置了已知密码的加密 PDF，系统亦具备解锁与解析能力，实测解析成功率保持在 98%以上，基本覆盖了常规业务场景。

OCR 识别则更多体现出行业适配的特征。系统并未直接使用原生 Tesseract 模型，而是引入超过 10 万张资质证书与发票图片进行针对性训练，使文字识别准确率由原本的 85%提升至 92%。面对部分扫描质量较差、分辨率不足 300DPI 的材料，系统会先执行直方图均衡化等图像增强操作，以减少噪点与对比度失衡对识别结果的影响。

文本清洗的工作方式更接近“减法处理”。一方面，通过哈希算法快速剔除如投标人承诺函这类高度重复的模板化内容，降低冗余信息对后续比对过程的干扰；另一方面，依托包含 500 余个行业无效词的词表，过滤掉无实质语义的填充性表述。与此同时，系统建立了专业术语映射机制，将“建筑工程施工总承包一级”等冗

长表述统一转换为“建筑一级资质”，从而在语义层面实现标准化，为后续的相似度计算提供一致的输入基础。

4.2 智能语义相似度计算模块

该模块为系统核心，采用 BERT 基础模型结合行业微调的方式，实现投标文件文本深层对比：

1）模型微调：以 BERT-Base 为基础模型，使用包含 50 万余条历史投标文件文本的招投标行业语料库微调，语料库覆盖通信、金融、交通等多个行业；微调采用 Adam 优化器，学习率设置为 2e-5，迭代 10 轮后模型损失函数降至 0）08；

2）语义向量生成：将预处理后的文本按 512token 长度分段，输入微调后的 BERT 模型，提取[CLS]向量作为语义特征，向量维度为 768 维；

3）相似度计算：运用余弦相似度算法计算两份文件语义向量的相似值，公式如下：

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），将 0-1 区间的结果转换为 0-100 分的百分制分值；

4）风险阈值设定：通过 ROC 曲线分析 1000 余例真实围标串标案例数据，设定相似度≥90 分为高风险、80-89 分为中风险、低于 80 分为低风险，该阈值下模型召回率 92%、精确率 88%。

4.3 格式与关键词比对模块

该模块从非文本语义维度补充相似性识别，提升检测全面性：

1）格式属性对比：提取文件作者、设备名称、修改时间、MD5 值等元数据，制定判定规则：两份文件作者与设备名称完全一致，或修改时间差≤10 分钟且 MD5 值相似度≥95%，标记为格式可疑；

2）关键词对比：通过正则表达式提取注册电话、邮箱、法人姓名、项目负责人资质编号等核心信息，计算信息重合率；若两份文件注册电话与法人姓名完全一致，判定为企业信息关联可疑。

4.4 结果可视化与风险预警模块

系统完成数据采集、预处理、对比、输出全流程功能测试，结果如下：

多格式文件导入测试 200 例，通过 196 例，通过率 98%，未通过原因为 4 个加密 PDF 未提供密码；OCR 文字提取测试 100 例，通过 92 例，通过率 92%，未通过原因为 8 个低分辨率模糊图片识别误差；

语义相似度对比测试 300 例，通过 258 例，通过率 86%，未通过原因为 12 份文件含生僻专业术语；

风险预警测试 50 例，通过 46 例，通过率 92%，未通过原因为 4 个低风险文件误判为中风险。

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；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	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

系统完成数据采集、预处理、对比、输出全流程功能测试，结果如下：

多格式文件导入测试 200 例，通过 196 例，通过率 98%，未通过原因为 4 个加密 PDF 未提供密码；

OCR 文字提取测试 100 例，通过 92 例，通过率 92%，未通过原因为 8 个低分辨率模糊图片识别误差；

语义相似度对比测试 300 例，通过 258 例，通过率 86%，未通过原因为 12 份文件含生僻专业术语；

风险预警测试 50 例，通过 46 例，通过率 92%，未通过原因为 4 个低风险文件误判为中风险。

5.3 性能测试结果

1）耗时指标：测试组 1 耗时 2.3 分钟、测试组 2 耗时 4.8 分钟、测试组 3 耗时 7.5 分钟，均符合预期要求，且耗时随投标单位数量线性增长，满足扩展性需求。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）准确率指标：系统语义相似度识别平均准确率 86%，高于 85%的需求标准；围标串标可疑案例识别率 92%，较人工 60%的识别率提升 32 个百分点；

3）并发指标：模拟 10 名用户同步上传比对，系统 QPS 达 200，页面响应时间 0.8 秒，满足性能要求。

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系统，通过分层架构、多维度对比模型、大数据预处理流程，实现了投标文件相似性的自动化、高精度检测。测试数据表明，系统处理 50 家投标单位项目仅需 7.5 分钟，较人工核查缩短 99.93%；语义识别准确率 86%，围标串标识别率 92%；单项目人工成本降低 95%，远超预期降本目标。该系统有效解决了传统招投标核查效率低、精度差、成本高的问题，为招投标合规管理提供了可靠技术支撑。

6.2 未来展望

后续将从三方面优化升级系统：一是优化算法模型，引入 GPT-4 轻量版本模型，提升生僻专业术语的语义识别能力，目标将准确率提升至 90%以上；二是拓展功能模块，新增投标报价规律分析功能，通过大数据分析报价等差、规律性变化等异常情况，强化围标串标甄别能力；三是适配更多场景，开发移动端应用程序，支持核查人员现场查看检测报告，满足外出核查工作需求。

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]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.

作者简介：（1）许雅思（1988-），男，汉，广东汕尾人，本科，工程师，任项目总监、专家；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。

（2）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。

（3）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。

（4）刘萍（1986-），女，回，广西永福人，本科，工程师，任项目总监、电气自动化及信息化领域专家；从事机电、电气工程及信息化项目开发管理 16 年，主要研究方向为机电和电力工程建设项目全过程管理及信息系统开发研究。

（5）叶明川（1987-），男，汉，广西桂林人，硕士研究生，工程师，任项目总监、电气自动化及信息化领域专家；从事机电、电气工程及信息化项目开发管理 11 年，主要研究方向为机电和电力工程、信息化建设项目全过程管理，以及 IT 系统开发研究等。

（6）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）。