计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(2):78−87 [doi: 10.15888/j.cnki.csa.008412]

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

非结构化智能金融投研平台的开发与行业应用①

陈　强

(兴业银行 信息科技部, 上海 201201)
通信作者: 陈　强, E-mail: aoyu_77@qq.com

摘　要: 当前针对非结构化数据处理的研究多集中于实验态的技术实现, 对于其在金融投研业务中落地应用的整体架构与路径的研讨则较为缺乏. 为此, 提出将大数据、自然语言处理、知识图谱等技术结合起来进行智能化投研平台的研发设计, 并实现其在真实金融投研场景的应用. 该平台基于 Hadoop 分布式系统进行数据采集、存储与计算, 集成了传统文本处理技术及主流 AI 算法, 形成了投研相关的深度语义理解能力, 一是高效提取出金融文本信息, 并以知识图谱的形式存储; 二是基此进一步挖掘预测, 输出金融投研领域的智能化分析服务. 以金融研究中城投债相关文本作为测试样例, 验证了平台运行效果, 结果表明平台能以较高的准确率全流程、自动化地实现各项功能, 提升金融投研领域的工作效率.

关键词: 智能金融; 非结构化投研平台; 语义理解; 知识图谱; AI 算法; 行业应用

引用格式: 陈强.非结构化智能金融投研平台的开发与行业应用.计算机系统应用,2022,31(2):78–87. http://www.c-s-a.org.cn/1003-3254/8412.html

Development and Business Application of Unstructured Intelligent Financial Investment-research Platform

CHEN Qiang

(Department of Information and Technology, Industrial Bank Co. Ltd., Shanghai 201201, China)

Abstract: Current research on unstructured data processing focuses on technological realization in experiments, while the discussions on the overall architecture and the practical application path in financial investment-research businesses are far from enough. Considering this, this study proposes to develop and design the intelligent investment-research platform through the combination of big data, natural language processing (NLP), and knowledge graph (KG) and implements the platform in real financial investment-research scenarios. With the data collecting, storage, and computation operated through the distributed system Hadoop, the platform integrates traditional text processing technologies and mainstream AI algorithms to form a deep semantic understanding capacity. As a result, the platform is capable of extracting financial text efficiently and storing the information in the form of KGs, and it can also provide intelligent analysis in financial investment research by further exploration and prediction. Taking the tests related to municipal bonds as samples, this research has proved the validity of this platform. The results show that the platform can automatically perform various functions with high precision through the whole process, thus promoting the working efficiency in financial investment research.

Key words: intelligent finance; unstructured investment-research platform; semantic understanding; knowledge graph (KG); AI algorithm; business application

① 基金项目: 2019 年国家互联网数据中心产业技术创新发展大会技术创新二等奖 (NIISA) 项目; 2019 年中国外汇交易中心银行间市场金融科技创新大赛三等奖项目; 2018 年上海市人工智能创新发展专项 (XX-RGZN-01-18-9814)
收稿时间: 2021-06-22; 修改时间: 2021-07-20; 采用时间: 2021-07-27; csa 在线出版时间: 2022-01-17

随着金融行业的数字化程度不断提升, 行业迅猛发展的同时也产生了海量的半结构化、非结构化等形态的数据, 这诸多类型的数据信息往往蕴含了丰富的金融业务知识与逻辑, 同时也对经济社会, 乃至金融业务的发展产生了越来越重要的推动作用. 对于金融投资与研究来说, 其核心在于能够从市场包罗万象的数据信息对资产未来的价格走势进行预测判别, 通过对信息的深入分析, 可以缓解信息不对称, 有助于实现更精准的投资决策. 在传统的金融投研工作过程中, 各类结构化数据往往是常用的主要信息来源, 其价值已在较大程度上得到释放; 而文本类信息由于在解析上有更大挑战, 其业务价值尚未能像结构化数据一样被充分利用, 需要通过新技术进行深度挖掘.

金融科技蓬勃发展, 已成为推动金融业务发展的新引擎[1]. 在大数据、人工智能、知识图谱等金融科技核心技术的驱动下, 投资研究也正在向智能化转变[2], 一方面通过自然语言处理等技术对非结构化数据的工程化处理, 能提升研究中数据采集、信息挖掘的效率与及时性, 并对跨渠道、跨领域的不同信息进行关联整合, 形成更丰富的知识体系; 另一方面通过机器学习算法模型形成更优的投资策略, 能增强投资中分析预测、趋势研判的精准性与前瞻性, 也能降低人为情绪波动对科学决策的影响. 为此, 打造以前沿技术为支撑的智能化投研平台, 是当前形势下资产管理业务提升投研效率、增强市场竞争力的新利器[3].

1 金融行业相关智能化投研建设与技术应用

1.1 国内外智能化投研平台探索与实践

国外较早涌现了各类智能化投研平台, 其应用主要是集中在非结构化数据的采集、抽取、整合、分析等方面, 为市场研判、投资决策等提供更深度的信息支持[4], 如 Kensho 的智能投研平台基于 AI 平台对经济社会各领域信息进行挖掘提炼, 并结合知识图谱技术构建了金融事件图谱, 能及时预测各类事件对金融资产价格的影响[5]; Alphasense 的智能投研平台通过自然语言处理与知识图谱技术对各种金融文档进行结构化、实体化、知识化沉淀, 形成对广泛金融信息进行交互式搜索、管理和再加工的服务. 国内的基金、证券等金融机构也在积极探索智能化投研平台的建设, 如天弘基金采用垂直搜索、网络爬虫、人工智能等技术打造了智能投研平台[6], 能及时捕捉金融市场信息, 并对各类金融事件进行关联分析; 工银瑞信基金智能投研平台基于自然语言处理与 AI 技术, 主要实现了数据抽取与解析、知识图谱、智能搜索、智能推荐等服务, 在对外部行业信息与内部研究成果进行整合、沉淀, 提升研究的价值.

1.2 智能金融投研平台核心技术应用研究

智能投研平台主要涵盖了信息提取、关系识别、情感分类等主要功能, 自然语言处理、AI 与知识图谱则是实现这些功能的核心技术. 罗平[7] 指出, 以 LSTM 为主的深度学习算法已成为金融文本信息提取的主流, 尤其是 Bi-LSTM 在命名实体识别、关系抽取等方面都表现出较大的优越性; 黄胜等[8] 采用 Bi-LSTM+CRF 神经网络结构, 并结合领域词典, 抽取金融文本中的结构化信息, 发现该方法较传统的规则匹配有较大提升效果, 且能满足多种类文本的提取需求; 陈剑南等[9] 采用双向 LSTM 结合多重注意力机制模型提取金融事件中实体间的关系, 并基于 Neo4j 数据库构建了金融事件图谱, 对金融事件之间的联系形成更精准的画像; 赵亚南等[10] 指出 Attention 机制已广泛用于各类文本任务, 并通过实验表明基于多头注意力机制的 Transformer 模型在金融文本极性分析上取得较好的效果; 马远等[11] 在目标方面词的左右分别采用 Bi-GRU 和 Attention 机制提取双边的语义信息, 并将双边特征与目标词结合起来识别文本的情感类别, 实现对文本语义更细粒度的处理; 赵澄等[12] 在对金融文本进行情感分类的基础上, 将情感的正负面类别作为关键特征加入股票预测模型, 实验结果表明该方法提升了股票价格预测的准确性.

智能投研相关的行业性研究主要集中在两方面, 一是侧重于行业应用与案例的介绍, 对智能化技术及平台构建方法的探讨尚有不足; 二是注重对实现某一项功能的技术探索, 缺乏对平台整体技术与应用架构的研究. 为此, 本文从智能投研平台建设的整体架构出发, 提出了融合大数据、自然语言处理、机器学习与知识图谱等技术的智能化平台的研发设计与应用实现方案, 重在探索智能化技术在金融投研场景中的落地路径及应用范式.

2 智能金融投研平台的开发设计

智能金融投研平台的建设主要由数据处理层、智能分析层、业务领域应用等部分构成, 对非结构化文

本信息进行解析、提取、整合与关联, 实现投资研究中知识搜索、分析预测等功能, 为金融投资、财富管理等各业务领域提供多维度的智能化投研服务. 该智能金融平台整体架构如图 1 所示.

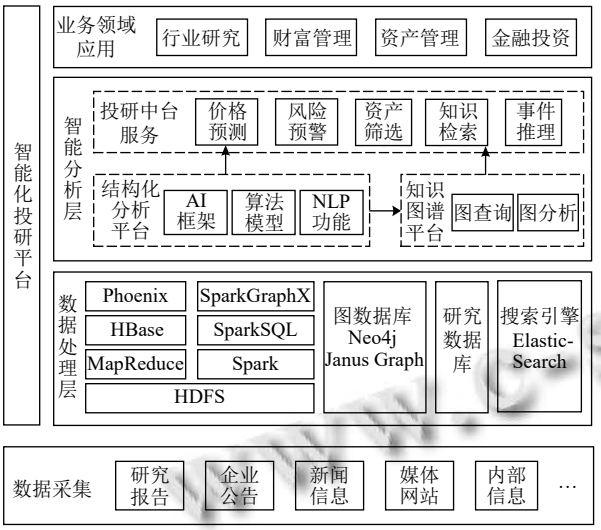

图 1 智能金融投研平台整体架构

整个平台最底层对接多种数据源, 涵盖了研究报告、企业公告、新闻信息、媒体网站等外部信息, 以及内部的投研相关信息, 采用 Sqoop/Flume 将各类数据源中的关系型及非关系型数据导入数据处理层中. 数据处理层集成了多种主流的大数据功能组件, 为多源异构数据的存储、转换、加工及复杂计算等提供能力支持. 该层采用 Hadoop 分布式系统处理架构, 通过 HDFS 实现对投研领域长时间、跨渠道数据的存储; 基于 HBase 和 Phoenix 解决对金融信息准实时查询的性能问题; 依托 SparkSQL 实现对行业信息的复杂规则计算及灵活的数据探索; 通过 SparkGraphX 图计算引擎进行全局的、多层次的金融知识图谱分析计算. 数据处理层还配置了 Neo4j、Janus Graph 等图数库, 支持从文本中提取出的数据信息以图结构的形式进行存储及查询展示; 同时, 这些预处理信息也通过构建索引的形式存储在 ElasticSearch 搜索分析引擎中, 实现高灵活性、高准确性、低延时及大规模并行化的检索查询[13].

智能分析层在数据处理层提供的存储、计算等资源支持下, 主要涵盖 3 部分: 一是将 AI 算法与语义理解领域技术相结合, 构建出面向投研领域的各类应用型算法模型, 形成对金融文本的多种处理能力, 从而提取出有价值的数据信息并进行相应的业务预测; 二是基于从金融文本中提取出的结构化信息, 构建金融知识图谱, 形成对金融知识与研究成果的沉淀与关联, 并实现快速的检索查询及分析推理; 三是整合语义理解及金融知识图谱的分析预测结果, 最终形成舆情分析、观点提取等多种投研领域的智能化中台服务, 可供行业研究、资产管理等不同业务领域灵活调用, 赋能金融业务的数字化、智能化转型.

2.1 非结构化分析子平台

结构化分析子平台是整个智能化投研建设的关键部分, 以自然语言处理技术和 AI 算法为核心支撑, 通过两者的结合, 尤其是深度学习算法的应用, 使平台进一步形成了对金融文本语义的深度理解能力, 从而更精准地实现各类文本非结构化任务以及智能化金融分析服务. 平台的这部分功能[14] 主要在于, 一是金融信息的工程化处理, 即对金融文本中的段落、句子、词语等进行细致地识别与提取, 并形成金融信息网络, 促进文本等非结构化信息的结构化、实体化以及标准化; 二是金融分析预测, 对各种纷繁信息进行灵活的再组合、再挖掘, 评估相关信息对金融事件或金融资产的影响. 该子平台的整体架构如图 2 所示.

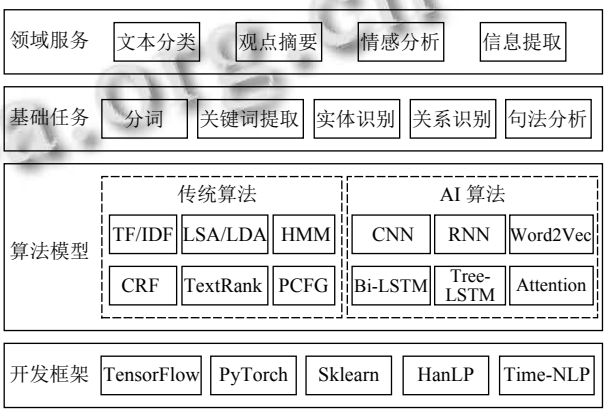

图 2 非结构化分析子平台

该子平台集成了各类 AI 基础算法及主流开发框架, 塑造了多维的语义理解能力, 形成了针对金融文本的分词、命名实体识别等基础任务, 最终构建了面向不同应用的金融文本相关分析服务. 在开发框架上, 除了 Sklearn、TensorFlow、PyTorch 等机器学习、深度学习框架, 还引入了 HanLP、Time-NLP 自然语言处理

等自然语言处理的专业框架, 以提升对文本进行基础处理的能力和效率. 基于丰富的开发框架, 平台涵盖了传统的自然语言处理算法及深度学习 AI 算法, 形成了更完善的基础算法体系, 能根据不同的金融文本非结构化信息的处理及分析目标, 对各类基础算法进行灵活的组合、重构与优化, 从而构建出相应的应用型算法模型, 也形成了针对金融文本非结构化信息的多维语义理解能力. 在此基础上, 平台一方面能高效地实现分词、关键词抽取、命名实体识别等处理任务, 为后续的金融分析服务提供更高质量的数据基础. 另一方面在结构化信息的基础上, 形成了对金融文本的信息分类、观点摘要、舆情分析等智能化应用服务.

结合工程实践与应用场景特点, 经多次反复验证模型效果, 本文确定了相关金融业务场景的模型技术架构. 考虑到金融文本处理相关任务及应用类型较多, 表 1 展示了本方案中最主要的几类任务.

表 1 金融文本结构化分析任务/服务 (示例)

任务/服务	功能	模型架构
实体识别	从文本中识别出特定的主体	BERT+CRF
关系抽取	判断不同主体之间的关联关系	WF+PF+CNN+Rankloss
情感分析	判断一段文字的正面或负面倾向	Bi-LSTM+Attention

命名实体识别是文本处理的一项基础任务, 主要目的在于从语料中提取出特定类型的主体, 是构建金融知识图谱的关键环节. 通用领域的命名实体大致分为人名、地名、机构名、时间、日期、货币和百分比等 7 类, 而在特定领域, 命名实体则有着更多更细致的类型, 如金融债券业务中, 债券名称、发行要素、经济指标、财务指标等均为重要的实体内容. 在实体识别的方法上, 时间、日期、货币、百分比等可以通过语言模板及正则表达式等模式匹配的方式被较好地提取出来, 而其他实体由于形式的多样化, 具有更高的提取难度, 需要借助 AI 算法模型实现更精准的识别. 本文方案中, 主要采用 BERT 与 CRF 结合的深度神经网络模型, 该模型在具体实现过程中所采用的技术架构如图 3 所示.

BERT 采用多层双向 Transformer 机制进行编码和解码, 能够将输入句子中的每一个词与其前后所有的词进行关联, 通过对长距离特征的捕获, 更完整地理解整体的语义信息[15]; CRF 将当前节点的输出序列与相连节点的输出序列相关联, 能更有效地解决序列标注和预测问题[16], 两者结合起来可以更精确地对文本进行划分, 并识别出语句的含义. 在命名实体识别过程中, 一般采用线性链 CRF, 当输入变量 X 取值为 x 时, 其输出变量 Y 取值为 y 的条件概率函数形式为:

$$P(y|x)=\frac{1}{Z(x)}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,t}\mu_l s_l(y_i,x,i)\right) \tag{1}$$

$$Z(x)=\sum_{y}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,t}\mu_l s_l(y_i,x,i)\right) \tag{2}$$

其中, t_k 和 S_l 分别是转移特征函数和状态特征函数, λ_k 和 μ_l 分别为对应的权重, $Z(x)$ 是规范化因子, 表示所有可能的输出序列的概率取值总和. 在实际金融业务建模中, 模型的训练目标为使真实序列发生的概率在所有可能生成的序列中占比最高.

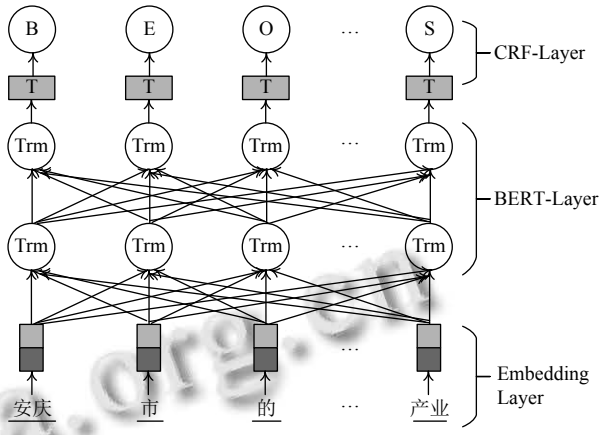

图 3 BERT+CRF 模型架构

关系抽取是指自动识别文本中实体对之间的语义关系类型, 是构建知识图谱、实现文本信息结构化的重要步骤, 也是文本处理的一项基础任务. 如, “任正非1987 年在深圳成立了华为公司”这句话, “任正非”相对于“华为”的关系为“创始人”. 目前关系抽取在完全监督、远程监督、联合抽取等不同方法上都有较多经典的算法模型, 本文方案中基于识别出的实体, 采用远程监督方法进行关系识别. 该模型的技术架构如图 4 所示.

该模型设计上, 先是构建词向量 (word embedding) 和位置向量 (position embedding), 在对词义进行刻画的同时, 也融入了实体对之间的位置信息; 然后采用

PCNN (piece-wise convolutional neural networks)[17], 根据实体的位置进行分段池化, 提取得到句子级的特征, 再通过 attention 对句子特征赋予不同的权重, 降低 instance 的噪声, 加权后的结果将形成整个 bag 的表征. 模型最后的输出为 bag 与每个 relation 的相似度, 作为在该 relation 维度上的得分, 此处用向量的点积进行相似度的计算[18], relation 的得分表示为:

$$f(b,c)=W^c\cdot s \tag{3}$$

其中, W^c 表示由每一类 relation 向量组成的关系矩阵; s 表示 attention 加权后整个 bag 的特征. 同时, 该模型将 pairwise ranking loss 作为优化目标[19], 在尽量增加正样本得分的同时, 尽量减小负样本的得分, 使正负样本之间形成更加清晰的区分, 构建的损失函数如下所示:

$$\begin{aligned}L=&\ln[1+\exp(r(\rho^+-f(b,c^+)))]\\&+\ln[1+\exp(r(\rho^-+f(b,c^-)))]\end{aligned} \tag{4}$$

其中, r 为缩放调整因子, ρ^+, ρ^- 分别表示正负样本的 margin, 表示标签为正样本的得分, 则表示负样本的得分.

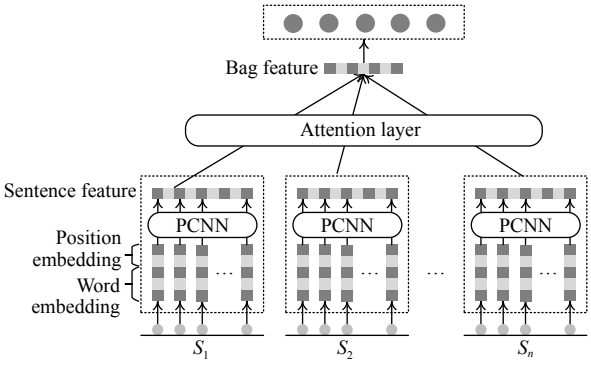

图 4 关系抽取模型架构

情感分析是指基于对一段文本的语义理解, 识别其所带有的情感色彩, 如正面或者负面等, 目前已具有较广泛的应用. 情感分析最常应用于分析客户对商品的评价, 以判断客户的满意程度. 在金融领域, 情感分析主要用于对新闻政策、公众舆论、社会事件等方面的挖掘预测, 捕获这些因素对企业或金融资产价格可能的影响, 以辅助进一步的分析决策. 在情感分析模型的构建上, 对正面、负面或者中性等的判断可以看成是一个多分类问题. 由于在一句文本中, 对于不同主体常常有不同的情感类型, 本文对经典 Bi-LSTM+Attention 模型进行改进, 针对目标词进行左右两边注意力机制提取关键特征, 形成针对特定主体的方面情感识别. 该模型技术架构如图 5 所示.

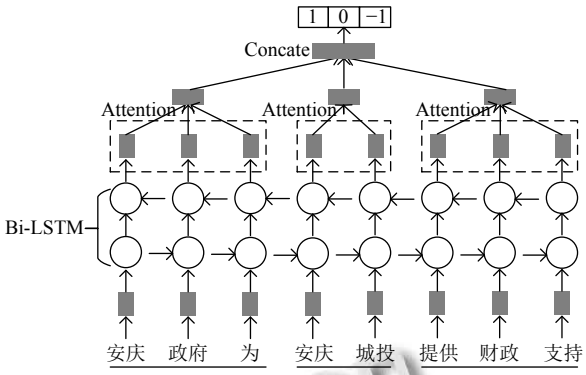

图 5 情感分析模型架构

以图 5 为例, “安庆政府为安庆城投提供财政支持”这句话中, “安庆城投”分析的目标主体, 以此确定左侧文本为“安庆政府为”, 右侧文本为“提供财政支持”. Bi-LSTM 在单向 LSTM 的基础上再增加一层反向循环层, 将正反两层网络的处理值拼接起来, 能全面地提炼出整条文本的语义特征[20]; 在此基础之上, 根据目标词的位置, 对左侧文本、目标词、右侧文本分别采用 Attention 机制进行特征赋权, 最终将 3 部分的特征组合起来形成针对目标词的整条文本表征向量 r_s, 具体计算表示如下:

$$r_s=H_1\alpha_1^{\mathrm{T}}+H_2\alpha_2^{\mathrm{T}}+H_3\alpha_3^{\mathrm{T}} \tag{5}$$

其中, H_i 为 Bi-LSTM 层在左侧文本、目标词、右侧文本这 3 个模块的输出, α_i^{T} 相当于 3 个模块 attention 层的权重矩阵; 最后再通过 *Softmax* 函数输出类别的概率, 给定输入文本 S, 其输出为:

$$P(y|S)=Softmax(W\cdot r_s+b) \tag{6}$$

采用交叉熵损失函数作为优化目标, 计算如下所示:

$$L=-\sum_{i=1}^{c}y_i\ln P_i \tag{7}$$

其中, y 表示样本的类别标签, $P(y)$ 表示模型预测样本为相应类别概率, c 表示类别的数目.

2.2 金融知识图谱子平台

金融知识图谱主要是通过大规模语义网络, 将金融领域中结构化、半结构化、非结构化等不同类型的数据进行整合. 图谱以实体或者概念作为节点, 节点之间以关系为边相连接, 通过图数据库以网络连接的形式进行可视化查询分析, 能够推动决策支持、个性化推荐等服务的智慧化发展[21].

在金融投研领域, 行业信息及研究成果等大都以

文本的形式存在, 信息之间缺乏关联性, 且分布零散, 难以形成对知识的沉淀. 通过知识图谱技术将跨领域、跨行业、跨主体的金融业务信息关联起来, 形成深度的金融信息网络, 对增强研究的深度与广度、提升金融投研工作的精准性有重要的价值和意义, 也能使搜索推荐、分析预测、查询决策等金融服务更加智慧, 进而增强金融机构投研业务的智能化水平. 本方案中, 金融知识图谱子平台的架构如图 6 所示.

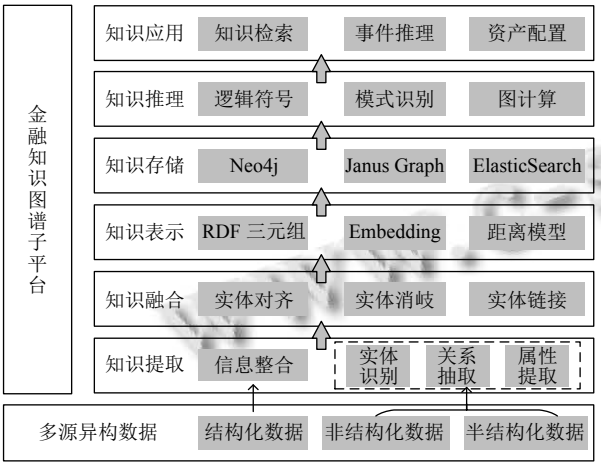

图 6 金融知识图谱子平台

金融知识图谱的构建基于结构化、非结构化、半结构化等不同类型的数据, 其中结构化数据, 如产能、产量等标准化数据可直接通过表结构转换进行信息整合提取; 而对于大量的非结构化金融文本来说, 需要依托前述的文本处理相关技术提取出文本中的实体、关系、属性等价值信息, 以进行后续的关联与分析, 这也是整个金融知识图谱开发中极为关键也具有难度的环节. 对提取出的信息进行同义实体的对齐、统一, 以及结构转换与信息关联, 并以“实体-关系-实体”“实体-关系-属性”的三元组等形式进行知识表示. 将融合好的知识信息存储在 Neo4j、Janus Graph 等图数据库, 以及 ElasticSearch 等查询引擎中, 以实现快速的图可视化查询, 以及深度的图挖掘分析, 最终支持知识检索、事件推理、资产配置等金融投研领域的业务应用[22].

3 智能金融投研平台的行业应用

本文以地方政府城投债相关金融文本作为实验数据, 依托 AI 算法, 基于自然语言理解能力对文本信息进行抽取, 实现金融文本的工程化、结构化处理, 并结合知识图谱技术进行实体、关系、属性的链接, 构建了城投债领域的金融知识图谱, 形成了债券金融信息网络. 最后对智能金融投研平台的信息抽取、知识检索、情感分析等主要功能进行测试, 实验结果表明该平台能实现从非结构化数据处理到知识检索、情感分析等各项功能的全流程、自动化、高精度运行.

3.1 语义关系

在进行金融信息抽取及知识图谱构建之前, 先基于城投债领域相关内容及业务逻辑进行语义关系的设计, 即确定城投债领域中实体的类型, 以及实体之间可能存在的关系类型, 这是进行后续开发的基础. 在具体实验中, 除了地名、机构名、时间、日期、货币和百分比等基本实体类型外, 还新增了行业、产业、债券名、经济指标、财务指标、经营业务、项目等领域特定实体. 在关系类型上, 重点搭建了涵盖空间、时间、物理、上下位等维度的语义关系架构, 图 7 展示了语义关系部分示例.

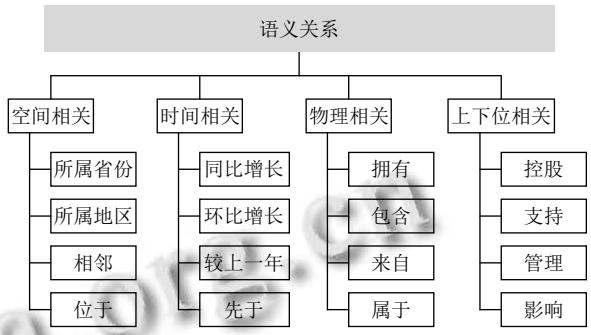

图 7 语义关系结构 (示例)

3.2 金融信息抽取

通过内外部渠道采集近 5 年内城投债行业的债券募集说明书、债券评级报告等金融文本, 共 10 000 篇左右, 作为平台实验的数据.

先基于 Jieba 分词器进行分词, 并通过添加城投债行业的词汇对词典进行维护, 提升分词的准确性; 在分词的基础上, 采用正则表达式与 BERT+CRF 模型相结合的方法进行相关实体的识别与抽取. 对于日期、时间、数值等这类形式较固定的信息, 直接采用正则表达式进行抽取; 而对于行业、公司名等内容较多样的实体, 正则表达式难以涵盖, 需采用前述的 BERT+CRF 模型进行识别. 图 8 为时间抽取的正则表达式示例, 通

过该正则表达式可以将时间内容, 以及对应的年、月、日等信息分别提取出来.

```
extract_date=re.match(
r'([0-9零一二三四五六七八九十]+年)?([0-9一二三四五六七八九十]+月)?
([0-9一二三四五六七八九十]+[号日])?', text_info)
if extract_date.group[0] is not None:
res={ 'year': extract_date.group[1], 'month': extract_date.group[2],
'day': extract_date.group[3] }
```

图 8 时间抽取正则表达式 (示例)

图 9 为安庆市城投债评级报告中的一段文本, 基于正则表达式与 BERT+CRF 模型相结合的信息抽取方法, 识别出了时间、百分比、行业、地区、公司名、经济指标等实体类型, 以及对应的具体文本内容. 实体识别的结果如图 10 所示.

石油化工、轻纺、机械仍是安庆的三大支柱产业，拥有中国石化集团安庆石油化工总厂（简称“安庆石化”），华茂集团、安徽曙光化工集团等相关行业的骨干企业。2018 年，安庆市实现规模以上工业增加值同比增长 8.3%，三大工业主导产业增加值增长 5.6%；其中，装备制造业增长 13.6%，纺织服装业增长 15.6%，石油化工业下降 2.5%。

图 9 实体识别文本 (示例)

石油化工 Domain、轻纺 Domain、机械 Domain 仍是 安庆市 Location 的三大支柱产业, 拥有 中国石化集团 Company 安庆石油化工总厂 Company (简称“安庆石化 Company”)、华茂集团 Company、安徽曙光化工集团 Company 等相关行业的骨干企业。2018 年 Time, 安庆市 Location 实现规模以上 工业增加值 Eco-index 同比增长 8.3% Percent, 三大工业 Domain 主导 产业增加值 Eco-index 增长 5.6% Percent; 其中, 装备制造业 Industry 增长 13.6% Percent, 纺织服装业 Industry 增长 15.6% Percent, 石油化工业 Industry 下降 2.5% Percent。

图 10 实体识别结果 (示例)

在识别出实体的基础上, 采用前述的关系抽取模型进行不同实体间关系的判断, 图 11 为一段城投债金融文本示例, 对其进行关系抽取的结果如图 12 所示.

3.3 金融知识图谱构建

基于语义框架中定义的城投债领域相关实体以及关系类型, 结合从金融文本中提取出的信息, 形成点和边的数据集, 通过数据库对数据进行存储连接, 开发出成型的投研领域企业债券知识图谱. 例如, 对于“城投公司属于某一地区”这一信息, 可表示为“城投公司—所属地区—地名”三元组形式, 在知识图谱构建过程中, 城投公司实体数据示例如表 2 所示, 地名实体数据示例如表 3 所示, 城投公司和地名之间关系数据集示例如表 4 所示, 将实体及关系数据集导入图数据库中, 即形成相应的关联关系图.

安庆市经济和财政保持较快发展，安庆城投得到市政府在财政补贴等方面的有利支持，子公司华茂股份生产工艺较为先进，在纺织行业中具有一定的竞争力与品牌知名度。

图 11 关系抽取文本 (示例)

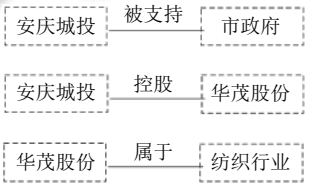

图 12 关系抽取结果 (示例)

表 2 公司名实体数据集 (示例)

company: ID	Label
安庆城投	company
沈阳公用	company

表 3 地名实体数据集 (示例)

location: ID	Label
安庆市	location
沈阳市	location

表 4 关系数据集 (示例)

START_ID	END_ID	RELATION	TYPE
安庆城投	安庆市	所属地区	所属地区
沈阳公用	沈阳市	所属地区	所属地区

将从债券行业文本中提取出的所有相关实体, 以及实体的关系和属性值导入图数据库中, 即得到较为完整的债券行业金融知识图谱 (如图 13 所示), 实现了金融文本的结构化、可视化, 有助于提升投研分析的便捷性与清晰度.

3.4 平台运行效果

以前述的实验设计为基础, 主要对报告的自动解析、知识检索及情感分析等方面的功能进行测试, 以

验证智能投研平台运行的效果及稳定性.

通过对债券相关文本中实体、关系、属性等金融信息的抽取、关联, 平台能够自动解析债券市场信息, 提取核心内容并进行金融知识图谱存储, 形成债券行业的知识库. 原本人工研读一份万字的债券评级报告需 2 h 左右, 而通过智能投研平台进行结构化处理仅需 1 min 30 s, 且解析效率高达 90% 左右, 能极大提升债券行业研究人员的文本阅读效率. 图 14 表示一份文本形式的评级报告, 图 15 表示智能投研平台对该评级报告进行解析后形成的结构化信息示例.

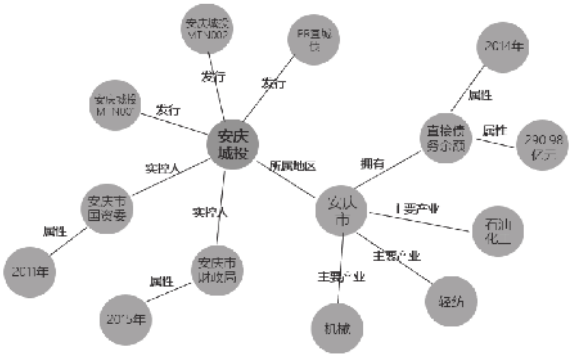

图 13 城投债关联图谱 (示例)


大公国际资信评估有限公司

公司债券跟踪评级报告

安庆市城市建设投资发展（集团）有限公司

2012 年公司债券跟踪评级报告

大公报 SD【2013】425 号

跟踪评级

公司债券信用等级：AA

主体信用等级：AA

评级展望：稳定

上次评级

公司债券信用等级：AA

主体信用等级：AA

评级展望：稳定

本期发债额度：18 亿元

上次评级时间：2012.7

债券存续期间：2012.12.31 – 2019.12.31

主要财务数据和指标

跟踪评级观点

安庆市城市建设投资发展（集团）有限公司（以下简称“安庆城投”或“公司”）主要从事安庆市的城市建设以及土地开发业务。评级结果反映了公司是安庆市重要的基础设施建设主体，继续得到地方政府的有力支持，2012 年以来安庆市经济继续保持较快发展、财政实力不断增强等有利因素；同时也反映了 2012 年以来安庆市全口径债务继续增长，土地出让收入仍是公司主营业务收入的唯一来源影响收入的稳定性等不利因素。

综合分析，大公对安庆城投 2012 年公司债券信

图 14 债券评级报告截图

解析出的结构化信息以知识图谱的形式进行存储, 形成金融文本信息的沉淀, 可供进一步的检索查询、推理预测. 人工阅读与智能投研平台解析的效果对比如表 5 所示.

基于图数据库存储沉淀的结构化信息, 并通过在 ElasticSearch 中构建倒排索引, 平台可以实现对金融信息快速、精确的查询检索. 图 16 与图 17 是信息检索示例, 在经过自然语言处理识别后, 图 16 中能够准确判断“GDP”与“地区生产总值”这类同义词, 图 17 中则能根据当年数据增长情况自动计算出上一年的数值.

图 15 报告自动化解析

表 5 智能解析效果

方式	所需时间	解析结果	知识应用
人工阅读	2 h左右	无标准形式	单次分析
智能解析	1 min 30 s	结构化数据	图谱存储、知识沉淀

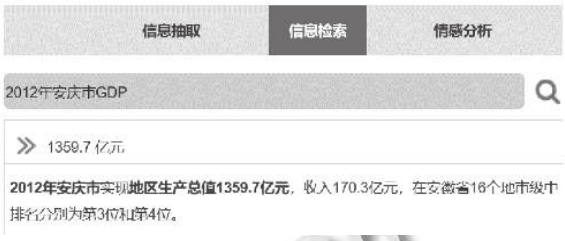

图 16 信息智能检索 (示例一)

图 17 信息智能检索 (示例二)

通过平台的智能化检索, 投研人员可迅速获得行业数据, 尤其是历史信息, 无须再从纷繁的文本报告中去查找, 极大提升了投研效率. 以一条历史信息的检索为例, 表 6 展示了人工检索与智能化检索的主要差异点.

表 6 智能检索效果

检索方式	形式	所需时间	检索内容
人工检索	大量文本查找	10 min左右	单一信息
智能检索	数据库自动搜索	几秒	关联展示

在对金融信息进行抽取、检索等基础工程化处理的同时, 基于对金融文本的分类预测, 平台还实现了对新闻报导、公司公告等信息的情感分析功能. 通过对信息的快速抓取与分析, 能够帮助投研人员及时、全面地了解市场动态, 以更前瞻地判断风险、更精准地判断金融资产价格的走势. 图 18 为一则关于安庆市城投公司的新闻, 图 19 为平台对这则新闻进行情感分析的结果, 针对“安庆城投”这一分析主体, “正面”表明新闻内容对其有利好影响, 相应的债券风险也降低. 大量文本的测试结果显示情感分析的准确率能达到 85% 以上, 对投研决策能起到较大的辅助作用.

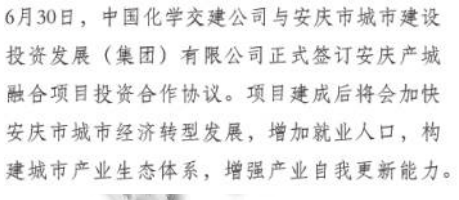

图 18 新闻信息文本 (示例)

图 19 平台情感分析结果示例

4 结论与展望

本文将大数据、自然语言处理、知识图谱等技术相结合, 提出了智能金融投研平台建设方案. 实验结果表明, 平台能以较高的准确率全流程、自动化地实现金融业务中相关命名实体识别、关系抽取、知识图谱构建等信息抽取与整合任务, 以及行业知识检索、情感分类等智能金融分析服务, 极大地降低了金融行业投研人员解析、查询金融信息的时间, 提升了投研工作的效率与精准度; 同时也实现了金融领域行业知识的多维度、持久化关联与沉淀, 为金融投资分析提供更加夯实的价值信息.

未来智能金融投研平台将进一步结合 AI 相关算法、知识图谱等技术领域的演进发展进行更深度的探索研发, 一方面持续提升在金融信息抽取、检索上的精准性; 另一方面积极探索事件推理、舆情因子等在金融资产配置及风险防控等方面的应用, 进一步提升平台在金融投研领域的服务价值.

参考文献

1 中国人民银行. 金融科技 (FinTech) 发展规划 (2019–2021 年). http://www.gov.cn/xinwen/2019-08/23/content_5423691.htm. [2019-08-23].
2 赵阳, 江雅文. 金融科技赋能证券经营机构财富管理转型研究. 金融纵横, 2019, (10): 36–45.
3 王超. 金融科技对商业银行资产管理业务的影响. 华北金融, 2019, (12): 51–57.
4 李嘉宝. 基于智能投研提高券商投研能力的探讨. 金融纵横, 2019, (6): 65–71.
5 倪隆洁, 田发. 浅析 Kensho 对我国互联网金融智能投研发展的启示. 经济研究导刊, 2020, (30): 61–62.
6 于天, 刘凯. 资管新规背景下人工智能在银行资管业务中的应用研究. 现代管理科学, 2019, (5): 25–27.
7 罗平. 金融文档语义理解——提升行业智能化的关键 AI 技术. 人工智能, 2018, (5): 94–104.
8 黄胜, 王博博, 朱菁. 基于文档结构与深度学习的金融公告信息抽取. 计算机工程与设计, 2020, 41(1): 115–121.
9 陈剑南, 杜军平, 薛哲, 等. 基于多重注意力的金融事件大数据精准画像. 计算机科学与探索, 2021, 15(7): 1237–1244.
10 赵亚南, 刘渊, 宋设. 融合多头自注意力机制的金融新闻极性分析. 计算机工程, 2020, 46(8): 85–92.
11 马远, 程春玲. 融合左右双边注意力机制的方面级别文本情感分析. 计算机应用研究, 2021, 38(6): 1753–1758.
12 赵澄, 叶耀威, 姚明海. 基于金融文本情感的股票波动预测. 计算机科学, 2020, 47(5): 79–83.
13 Li RZ, Li BJ, Zhang GZ, *et al*. A high-performance and flexible chemical structure & data search engine built on CouchDB & Elastic Search. Chinese Journal of Chemical Physics, 2018, 31(3): 341–349.
14 陈强, 代仕娅. 大数据、AI 平台支撑下的智慧金融产品研发与实践. 软件导刊, 2021, 20(2): 31–39.
15 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: ACL, 2019. 4171–4186.

16 田梓函, 李欣. 基于 BERT-CRF 模型的中文事件检测方法研究. 计算机工程与应用, 2021, 57(11): 135–139.
17 Zeng DJ, Liu K, Chen YB, *et al*. Distant supervision for relation extraction via piecewise convolutional neural networks. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015. 1753–1762.
18 Ye H, Chao WH, Luo ZC, *et al*. Jointly extracting relations with class ties via effective deep ranking. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: ACL, 2017. 1810–1820.
19 Santos CND, Xiang B, Zhou BW. Classifying relations by ranking with convolutional neural networks. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing: ACL, 2015. 626–634.
20 Zhou P, Shi W, Tian J, *et al*. Attention-based bidirectional long short-term memory networks for relation classification. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: ACL, 2016. 207–212.
21 黄梦醒, 李梦龙, 韩惠蕊. 基于电子病历的实体识别和知识图谱构建的研究. 计算机应用研究, 2019, 36(12): 3735–3739.
22 陈强, 代仕娅. 基于金融知识图谱的会计欺诈风险识别方法. 大数据, 2021, 7(3): 116–129.