

证券交易是一个历史悠久而又与时俱进的行业,从 17 世纪最早的证券交易所建立开始,证券交易就一直有着从纯粹的主观交易向着科学化理论化交易发展的趋势。20 世纪 60 年代到 80 年代,随着金融理论的发展和计算机的诞生,量化交易渐渐在金融领域得到广泛应用。



进入 21 世纪以来,随着人类社会与科技的进一步发展,证券交易所可以使用的工具越来越丰富,涉及的技术也越来越复杂而多样。人工智能技术正是这些技术中的重要一环。各种各样的人工智能技术为证券交易行业带来了新的变革。

时至今日,已有诸多知名私募建立了自己的 AI 研究部门或是 AI 策略开发部门,甚至部分私募曾对外宣称已将 AI 技术大规模应用于实盘,与 AI 相关的策略占比超过 50%。无论这是私募用于宣传的噱头还是确有其事都无法否认 AI 技术已经大范围进入证券交易从业者的视野,并有可能为这个行业带来一场革新。

AI 技术对于证券交易的帮助也显而易见:

(1) AI 能够通过强大的计算能力和复杂的算法分析海量的交易数据。它可以迅速地识别和发现隐藏在数据中的模式、趋势和关联性,帮助投资者做出更准确、更有依据的决策。与传统的人工分析相比,AI 能够更快、更全面地理解市场动态,并提供更可靠的预测能力。

(2) AI 在风险控制方面也发挥着重要作用。通过对历史交易数据的分析,AI 可以识别出潜在的风险因素,并进行实时监测和预警,以及时调整投资组合,最大程度地降低风险,保护投资者的利益。

(3) AI 有助于提高工作效率。以 ChatGPT 为代表的大模型已经在不知不觉中渗透进人们的工作生活。除了常用的代码提示以外,善用自然语言大模型还可以有效地提高人们的信息获取效率。

AI 技术应用于证券交易已经成为不可忽视的趋势。它通过深度学习、传统机器学习和大数据分析等技术手段,为投资者提供了更准确、更高效、更可靠的信息获取,以及交易决策支持和风险控制。

本章的主要内容包括证券交易的发展历程,AI 时代的量化策略开发与传统量化策略开发的异同及 AI 技术在实际量化开发场景中的应用。

3.1 证券交易发展历程

证券交易自诞生以来,经历了主观投资、理论投资、量化投资、大数据驱动和智能投资 5 个主要阶段,如图 3-1 所示。

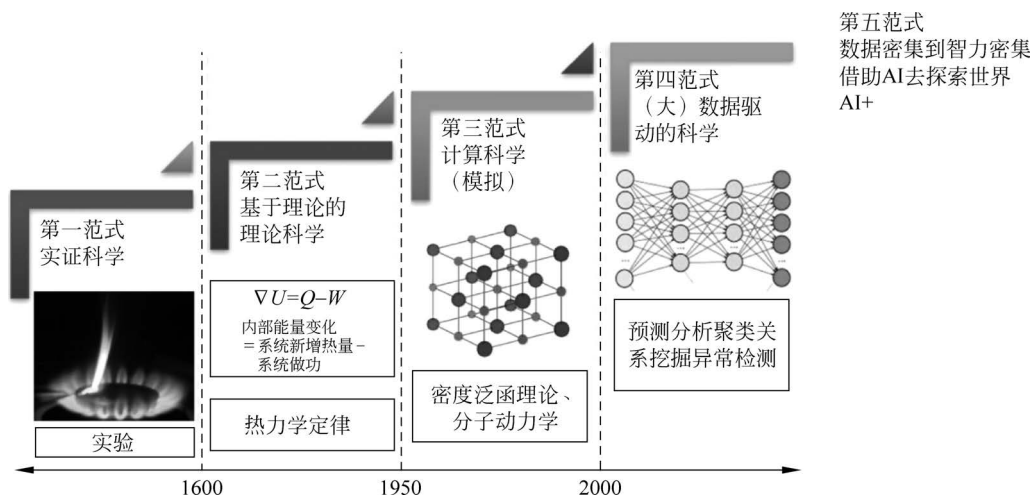


图 3-1 证券交易发展历程(图片来源于 BigQuant)

以下是对主观投资、理论投资、量化投资、大数据驱动和智能投资各阶段的历史和变迁的简要概述,以及一些主要代表人物。

(1) 主观投资：主观投资是指基于投资者主观判断和经验的投资方法。这是早期投资方法的主流,依赖于投资者的直觉和分析能力。

历史与变迁：主观投资一直是投资领域的主要方法,直到 20 世纪中叶。投资者如巴菲特、彼得·林奇和乔治·索罗斯等通过自己的洞察力和经验取得了巨大成功。

(2) 理论投资：理论投资是指基于经济学和金融理论的投资方法。投资者利用经济学和金融学的原理和模型进行投资决策。

历史与变迁：20 世纪 50 年代到 70 年代,随着现代金融理论的发展,理论投资成为主流。马克·约瑟夫和哈里·马科维茨等的资产组合理论和有效市场假说为理论投资奠定了基础。

(3) 量化投资：量化投资是指利用数学模型和统计方法进行投资决策的方法。投资者使用大量数据、算法和计算机技术来执行交易策略。量化交易的优势在于纪律性、及时性、系统性、准确性和分散化。量化投资需要经历数据获取、策略回测、模拟交易和实盘交易等流程。回测是量化投资的重要环节,并且需要注意未来函数、过拟合、幸存者偏差、冲击成本和小样本视角等陷阱。评价量化策略常用的指标有夏普比率、不同的盈利和抗风险能力指标。回测结果的资金曲线可以用不同指标来评估策略的表现。

历史与变迁：20 世纪 80 年代以来,随着计算机技术和数据处理能力的提升,量化投资迅

速崛起。代表性的量化投资公司包括文艺复兴科技、DE Shaw 和雷曼兄弟的 Alpha 策略。

(4) 大数据驱动：大数据为量化投资提供了海量信息、充足的样本量，以便于利用统计模型和机器学习对历史数据进行分析，预测股票、外汇等金融产品的未来价格走势；通过找出数据中的相似模式，将金融资产分类，发现不同类型资产之间的关联性，可以用于资产配置和组合优化；分析不同类型数据之间的关联关系，例如新闻事件和股票价格变动之间的关系，可以发现驱动市场的关键因素；通过识别数据中的异常点，发现例如股票市场中的操纵等异常交易行为，可以防范投资风险。大数据的驱动使投资策略更智能化和模型泛化更健壮化，有助于投资决策的准确性和交易效率的提升。

(5) 智能投资：智能投资是指利用人工智能和机器学习等技术进行投资决策和交易的方法。通过训练模型来自动化决策过程，以提高投资的准确性和效率。典型的特征就是从数据密集到智力密集，借助 AI 去探索世界，建立 AI 驱动的 Pipeline。人工智能的基本应用如图 3-2 所示。

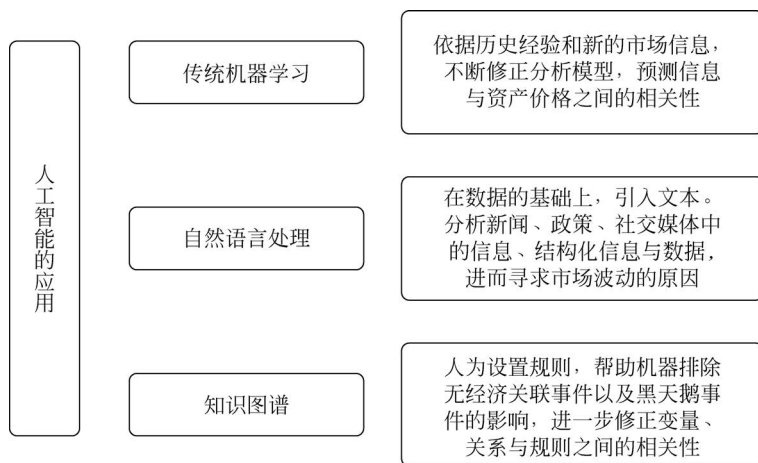


图 3-2 人工智能的基本应用(图片来源于知乎 collin)

历史与变迁：21 世纪以来，随着人工智能技术的发展，智能投资开始兴起。利用机器学习、深度学习和自然语言处理等技术，投资者可以分析大规模数据、发现模式并生成交易信号。代表性的智能投资公司包括 Two Sigma、Bridgewater Associates 和 Renaissance Technologies 的 Medallion 基金。需要注意的是，这些投资方法并非相互排斥，而是随着时间的推移和技术的进步逐渐发展和演变的。许多投资者和公司在实践中会结合不同的方法和技术，以适应不同的市场环境和投资目标。

3.2 AI 时代的量化策略开发与传统量化策略开发比较

3.2.1 传统策略开发的问题

传统策略的开发如图 3-3 所示。

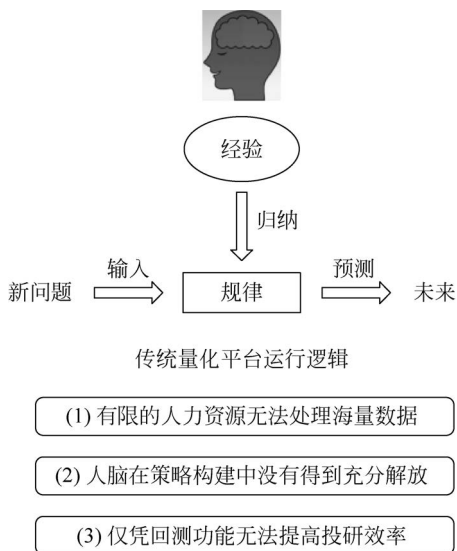


图 3-3 传统策略的开发(图片来源于网络)

(1) 传统量化策略的开发高度依赖经验积累,策略开发周期较长。

在投资领域,一个优秀的投资者能够获得正收益的关键在于他多年来积累的丰富投资经验。通过对这些经验的复用,他可以对未来的情况做出有效的决策。从机器的角度来看,这些经验可以被理解为一系列的数据。人类的精力是有限的,在这个大数据时代,机器可以获取的数据远远超过人类,因此,机器拥有比人类更加丰富的经验。普通投资者可能需要五年甚至十年才能积累到的经验,对机器而言只需几分钟的时间。机器可以利用相关算法从历史海量数据中找出潜在的规律,并生成预测模型,以指导投资者进行投资决策。此外,情绪是投资的一大障碍,而机器能够消除情绪的干扰,通过客观和理性的数据分析进行决策。可以预见,在大数据时

代,人工智能在很多场景下能够比人类更好地进行投资决策,也有助于普通投资者在短时间内赶超经验丰富的投资者。

传统的量化策略依赖于行情数据、财务数据和基本面数据进行研究和开发,它依赖于交易员和研究人员的经验和背景知识,对计算资源的要求不高。同样由于对经验的依赖性,传统量化策略在开发过程中,常常面临的一个现实问题是,当一些经典的策略或者因子被挖掘完后,形成一个具备较强创新性的可用的交易策略需要漫长的周期,并且策略验证也需要较长时间。

(2) 传统量化策略容易同质化。

尽管传统量化策略在回测时表现良好,但在实际交易中可能出现因子失效或市场风格切换而导致策略业绩回撤的情况。这也是传统量化投资的一个痛点,主要原因在于策略同质化的问题,即使用相同的数据和因子导致策略的同质化程度较高。此外,传统策略存在收益瓶颈,因为使用简单模型来捕捉因子和收益之间的线性关系在收益上存在限制。

3.2.2 AI 驱动量化策略开发的特点

AI 时代的量化策略开发相较于传统量化策略开发有以下特点。

(1) 更强大的数据处理和分析能力: 在 AI 时代,由于硬件计算能力的提升和更先进的硬件驱动、科学计算库的支持,计算机可以处理更大规模和复杂的数据集。传统量化策略在开发过程中,通常使用统计方法和基于规则的模型进行数据分析,而 AI 时代可以利用机器学习和深度学习等算法来发现更复杂的模式和关联。

(2) 更丰富的算法选择和模型构建: 在传统量化策略开发中,通常使用经典的统计方法算法,如线性回归、逻辑回归等,而在 AI 时代,可以利用更复杂的深度学习模型,如卷积

神经网络、Transformer 和图神经网络等,以此来构建更强大的预测和决策模型。

(3) 自动化和智能化: AI 时代的量化策略开发更加注重自动化和智能化。在传统量化策略开发中,许多任务需要手动完成,例如数据清洗、特征选择、参数优化等,而在 AI 时代,可以利用机器学习和自动化技术实现这些任务的自动化,减少人工干预并提高效率。

(4) 并行计算和云计算: AI 时代的量化策略开发更加注重并行计算和云计算的应用。传统量化策略开发通常在本地计算机上进行,受限于计算资源的限制,而在 AI 时代,可以利用分布式计算框架和云计算平台,以及高性能的计算实例,来加速策略开发和回测过程。

总体而言, AI 时代的量化策略开发利用了更强大的计算能力、更复杂的算法和更大规模的数据,以实现更精确的预测和更高效的策略。不过,即便如此,传统量化策略开发的经验和知识仍然具有重要价值,因为它们可以提供对市场和策略的深入理解,并为 AI 时代的量化策略开发提供指导和验证。

3.2.3 策略开发流程异同

策略开发流程如图 3-4 所示。

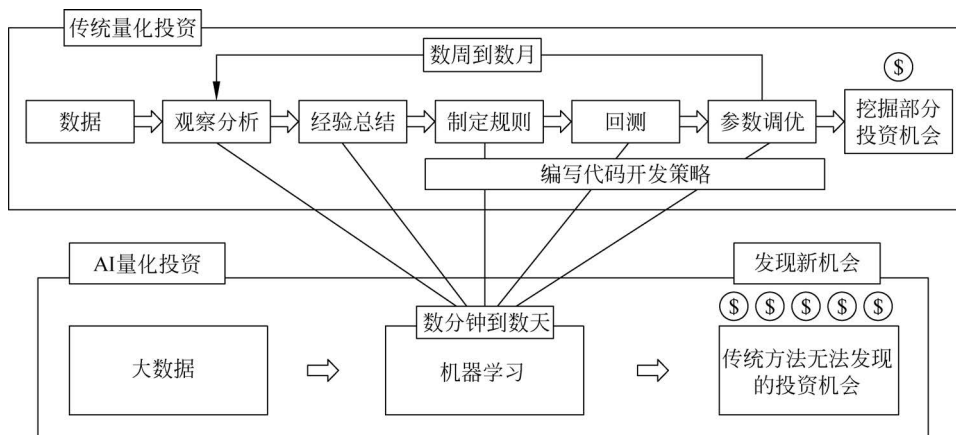


图 3-4 策略开发流程(图片来源于 BigQuant)

1. 传统量化策略开发流程各阶段特点

(1) 策略构想: 传统量化策略在开发过程中,策略基本构想的形成主要依托于研究人员对于市场细致的观察分析与历史经验总结,涉及对市场、资产类别或特定因素的研究和分析,以及对预期市场行为的假设。可解释性较强,相对开发周期也更长。

(2) 数据获取与清洗: 在传统量化策略的开发过程中,需要获取相关的市场数据、财务数据和基本面数据,数据量需求相对不严格。很多经典的统计模型对数据本身做出了较多的假设,例如假设自变量和因变量之间存在较强的线性相关性,这对数据本身的性质要求相对较高,因此在数据的选取和预处理时需要更加严谨。

(3) 因子构建: 高度依赖于量化研究员对市场的认知与灵感,高度依赖数理基础、统计基础。部分研究员可能会陷入随着入行时间增长,挖出有效因子的速度逐渐放缓的困境。

(4) 模型构建:传统量化策略在开发过程中,从业者会倾向于使用高稳定性、高可解释性、低复杂度的模型来对因子进行使用,例如线性回归、逻辑回归等。

(5) 参数调优:传统量化策略在开发过程中会使用一些较为常见的运筹优化算法,或者根据研究员的统计结果和主观理解来对策略的参数进行调优。

(6) 监控与修正:根据市场变化和策略绩效对策略做出相应的调整。

2. AI 量化策略开发流程各阶段特点

(1) 策略构想:理论上讲,只要数据与市场未来的走向有一定的关联性,数据的信噪比高于一定值,就能以这一部分数据为基础开始构建 AI 策略,对人工观察分析和经验总结的需求大大降低。

(2) 数据获取与准备:数据获取方面,如策略构想中所言,数据只要与市场未来的走向有一定的关联性即可纳入模型的考虑范围。数据处理方面,AI 量化策略开发在传统量化策略开发的基础上,有更多的 AI 工具可以用于缺失值、异常值的处理。相对而言,AI 策略对数据之间关联的形式要求更低,部分算法对于缺失值、异常值、类别特征数据自带了特殊处理,减少了数据处理方面的工作量。同时,AI 策略拥有更强的处理大规模数据的能力,将大量高质量的未经处理的原始数据直接输入模型,有时也能起到不错的效果。

(3) 因子构建:具有较高的自由度。AI 模型中的特征可以是另一个 AI 模型的输出,也可以选择不构建因子作为模型特征,直接使用原始数据或者原始数据经过简单处理后得到的特征直接进行预测。

(4) 模型构建:AI 策略在模型构建上具有广泛的选择空间,类似于 XGB、LGBM、CAT 的树模型,类似于 LSTM、GRU、Transformer 的时间序列模型,针对图结构数据的图神经网络等,或是某个单独的网络模块,甚至是某个网络的某一层都可以成为模型结构的一部分,AI 策略所使用的模型一直在跟随学术界最新研究成果的步伐。

(5) 参数调优:可以使用各类先进的 AI 工具进行自动化的参数调优,也可以使用 K-fold 等经典算法进行模型效果的验证和参数选择,核心思想是减少模型在数据集上的过拟合,使模型具有更强的泛化性能。

(6) 监控与修正:可以专门训练 AI 模型来应对各种市场突发情况,对异常情况的判断具有更高的敏感性和准确率,对异常情况的处理也更加灵活。

通过对比分析可以看到,AI 量化策略开发在许多方面具备传统量化策略开发所不具备的优势,如图 3-4 所示。尤其是使用 AI 进行量化策略开发会大大降低策略的生产周期,也降低了对研究员市场经验和数理基础的要求。这种开发方式能够在兼顾策略效益的同时,大大提升新策略开发速度。由此可见,AI 助力量化策略开发已经是不可阻挡的时代潮流。

不过,在利用 AI 进行量化策略开发的同时,需要认识到 AI 在目前阶段还仅仅是一种工具,神经网络较低的可解释性、对标签的依赖、部分模型较低的可控性、较强的过拟合风险都是 AI 量化策略在开发过程中可能面临的问题,一味相信 AI 也是不可取的。

推荐的做法是,可以先尝试着使用一些先进的 AI 方法,局部应用于传统量化策略在开发过程中的某些环节,经过市场和时间的检验后,再逐步地将各类 AI 融入策略开发的全流

程,提高策略开发效率。

3.3 AI 技术在量化开发场景下的应用

3.3.1 策略灵感来源

在大模型和大数据的加持下,研发量化策略的思路可以来自以下几个主要方面。

(1) 学术研究和文献:学术界和金融研究领域会经常涌现出新的理论和方法。量化策略开发者可以通过阅读学术期刊、金融论文和相关书籍来了解最新的研究成果和思路。

(2) 大数据分析:量化策略开发者可以利用大数据分析技术对海量市场数据产生整体的认知,通过统计分析工具和高效的数据可视化工具提取出其中有效的信息,并基于统计和分析的结果构建策略。

(3) 数据挖掘:利用数据挖掘技术,量化策略开发者可以从大规模数据中发现潜在的模式和规律,这些技术可以帮助识别趋势、寻找相关性和生成预测模型,为策略的开发提供有力支持。数据挖掘与大数据分析的主要区别在于,大数据分析一般会得到一个或多个统计量,如方差、均值等,这些统计量需要与业务结合进行解读,才能发挥出数据的价值与作用,而数据挖掘则更侧重于通过各类 AI 或者统计学模型直接获取知识。

(4) 交叉学科合作:量化策略开发涉及多个领域,如数学、计算机科学、经济学和金融学等。交叉学科合作可以带来更全面和创新的思路,例如金融学家、数据科学家和程序员之间的合作。

(5) 社区和论坛:在量化交易领域有许多在线社区和论坛,交易者可以在这些平台上与其他从业者交流经验、分享思路,并获取来自社区的反馈和建议。推荐使用语义搜索定位相关信息,以便快速获取所需知识。

(6) 自然语言处理(NLP):对于一些基于新闻和事件的策略,NLP 技术可以用来从新闻报道和社交媒体等非结构化数据中提取情感、舆论等因素,辅助策略的决策。推荐使用情感分析、观点抽取等算法。

量化研究员可以从以上来源得到新策略研发的启发和支持。与此同时,有条件的机构也可以建立私有知识库和语言模型,以便捕捉研发团队的语义信息,从而帮助新成员快速理解策略的设计思路和历史沉淀。高效利用 AI 工具,量化交易者可以开发出更加复杂和高效的策略,从而在市场中找到更多的投资机会。

3.3.2 策略解读与编码

量化开发者常常遇到的几种情况是:从某个会议或者期刊论文中发现了一篇感兴趣的论文后,由于各类因素的制约,无法快速地进行复现;在网络上发现一个公开代码仓库拥有自己想要学习的内容,但是由于代码风格等问题无法快速阅读;在某论坛上看到一个似乎有效的策略思路,由于语言问题或者表述风格问题,无法快速理解策略的具体内容,无法在短时间内深入策略的本质,无法判断策略是否合理。

以上情况在此前都是非常令人头痛的问题,但是自从以 ChatGPT 为首的自然语言大模型面市以来,此类困境得到了大大缓解。

一些自然语言大模型在训练的过程中,通过学习大量的源码、文档和编程语言规范,建立了对编程语言语法结构、变量、函数等编程概念的理解。当用户向模型提供一些与编程相关的提示、问题或示例时,它可以基于这些先前学习到的知识和模式生成相应的代码。当用户直接输入一段代码时,它也可以快速地对代码进行解读并添加注释,辅助用户对代码进行理解。

下面列举自然语言大模型在策略解读和编码中的部分应用场景:

- (1) 输入代码,使用模型解读代码,为代码添加注释,帮助快速理解代码。
- (2) 输入代码,使用模型为代码调试。
- (3) 输入想要实现的功能,使用模型寻找相应的库。
- (4) 输入代码框架,使用模型编写初版代码。
- (5) 输入某个代码仓库的文档,使用模型解读仓库的使用。
- (6) 输入策略文本和公式,使用模型分析其核心逻辑,并提供简单实现。

需要用户注意的是,自然语言模型的输出是根据上下文生成的,并没有从逻辑上对输出的代码或者知识点进行验证,因此,在使用模型生成的代码时,仍然需要人工审查和测试,以确保其正确性和可靠性。

同时,自然语言模型也受限于其预训练数据的质量和多样性。如果模型在预训练过程中没有充分接触到某些特定编程语言或领域的文本数据,则可能在生成相关代码时表现较差。相应地,如果用户所需的是比较常用的或者逻辑较为简单的代码,例如常见的 NumPy、Pandas 库的操作,则大模型的生成准确率会较高,使用大模型将会大大提高工作效率。

推荐的操作流程是,使用大模型生成所需的内容后,再进行一遍人工校审或试运行。

3.3.3 数据获取

使用 AI 工具可以辅助开发者获取数据,包括但不限于向开发者推荐数据源,帮助完成目标到代码的转义。

下面列举了 AI 在数据库抽取和生成 SQL 语句方面的具体应用案例。

1. 自然语言到 SQL 转换

使用自然语言处理技术,解析表达数据查询意图的自然语言,转换成 SQL 语句。例如用户询问“销量排名前 10 的产品有哪些”,可以转换为对应的 SQL 语句。

2. 图形界面 SQL 生成

通过图形化的查询拖曳界面,辅助用户生成 SQL 语句,降低 SQL 语言门槛。用户通过图形化操作指定查询逻辑,转换为 SQL 语句执行。

3. SQL 自动完成功能

IDE 可以根据数据库 Schema、表结构及用户已输入的部分 SQL 语句,使用机器学习技术推测用户意图,提供自动完成功能。

4. SQL 语句优化

分析查询计划,使用启发式算法或强化学习来重新规划查询执行计划,优化 SQL 语句,使其运行更高效。

5. SQL 语句检查

基于规则的方法或机器学习方法自动检查 SQL 语句,发现语法错误、安全漏洞、反模式等问题,给出修改建议。

6. SQL 工作负载管理和监控

使用机器学习技术分析 SQL 查询日志,进行工作负载管理,并主动监控 SQL 语句性能,防止突发问题影响在线服务。

7. 表关系抽取

从数据库或非结构化文档中抽取表名称、列名称、表关系等结构化信息,辅助程序理解数据库 Schema。

由此可见,AI 技术可以辅助开发者更便捷高效地利用 SQL 进行数据查询和管理。

BigQuant 平台利用 ChatGPT 专门训练了一个 AI 因子助手(X2),可以快捷地获取所需要的因子数据。通过该 AI 助手获取数据的案例如图 3-5 所示。

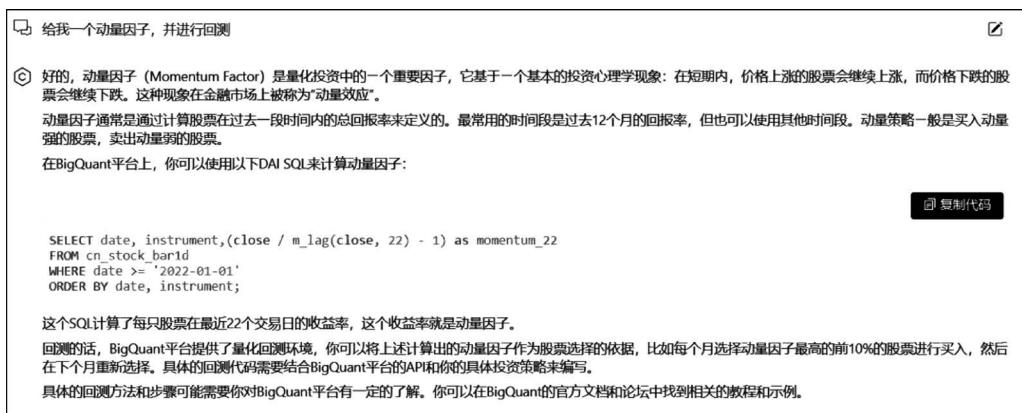


图 3-5 AI 助手获取数据的案例(图片来源于 BigQuant)

3.3.4 数据表征与模型构建

利用 AI 模型进行市场相关信息预测,或将 AI 模型的输出结果作为一种数据特征,是现在很多金融机构在构建量化策略中的常用手段。

深度学习作为一种表示学习的形式,擅长从高维、非结构化的数据中提取层次化的特征。它通过对大量简单但非线性的函数进行复合来学习这种层次化的表示,从输入开始,这些函数将一个层次的表示逐步转换为一个可能包含更抽象信息的新表示。经过足够多的非线性变换,深度学习能够拟合非常复杂的函数。

通过深度学习,可以跳过大部分手工设计信息性特征的工作,直接学习数据的内在结

构,更好地捕捉其与结果有关的特征。在理想情况下,这些特征能捕捉到结果的独立驱动因素,提高策略的收益。

下面将会简要介绍常见的深度学习模型在量化领域的应用。

卷积神经网络(CNN):该网络常用于视觉相关的机器学习任务,但根据该网络的原理,只要数据符合模型的关键假设,即局部模式或关系有助于预测结果,该网络就能在时间序列预测问题上起到一定效果。在时间序列的背景下,局部模式可以是相关间隔的自相关或类似的非线性关系。沿着第二和第三维度,局部模式意味着多变量序列的不同组成部分之间的系统关系,或不同股票的这些序列之间的关系。由于局部性很重要,因此如何组织数据是使用CNN进行时间序列预测的重要课题。

循环神经网络(RNN):RNN是一种经典的专门处理序列数据的网络结构,后续的LSTM、GRU、Transformer都受到了该网络的启发。这些时序网络有着千丝万缕的关联,后面的内容将不再一一赘述,只对此类网络的应用做一个笼统的概括。

相比于卷积神经网络,此类网络更加强调输入的时序性,其假设以前的数据点会影响到当前的观察,因此,它们不仅需要学习更加复杂的输入输出关系,也需要保留非常长远的信息,同时还需要对距离较远和距离较近的信息做出取舍。有非常多的金融标的是天然的时间序列数据来源,将此类网络用于价格预测是非常自然的选择。

生成对抗网络(GAN):生成对抗网络是生成模型家族的一分子。它的主要思想很简单,在一个竞争环境中训练两个神经网络,分别称为生成器和判别器。生成器的目标是产生判别器无法从给定的训练数据类别中区分的样本,而判别器的目标是判断数据是生成的还是真实的。最终的目标是获得一个生成模型,能够产生代表某种目标分布的合成样本。

生成对抗网络可以用于生成虚假的金融数据,辅助某些监督学习算法进行训练。也有一些方法尝试直接使用训练好的网络中的一些参数,将其视为金融数据的一部分特征,放入其他模型中进行市场价格的预测。

3.3.5 策略调优

在策略调优和参数搜索上常用的工具和应用方案包括以下几种。

(1) 强化学习(Reinforcement Learning):强化学习是一种通过与环境交互来学习最佳行为策略的机器学习方法。在策略调优中,强化学习可用于通过试错学习找到最优的策略。它可以根据奖励信号对不同的策略进行评估和优化,从而找到最优的策略参数。

(2) 遗传算法(Genetic Algorithms):遗传算法是一种模拟自然进化过程的优化算法。通过模拟基因的交叉、变异和选择,遗传算法可以搜索参数空间中的最优解。在策略调优和参数搜索中,遗传算法可以用来生成候选解并根据适应度函数对其进行排名和选择。

(3) 贝叶斯优化(Bayesian Optimization):贝叶斯优化是一种有效的全局优化方法,适用于黑盒函数和高维参数空间。它通过建立参数的概率模型,并在更新观测数据后不断调整该模型,以此来指导下一次参数采样的选择。贝叶斯优化在策略调优和参数搜索中可以用于高效地探索参数空间并找到最优解。

(4) 网格搜索和随机搜索(Grid Search and Random Search): 这是一种简单但常用的参数搜索方法。网格搜索通过指定参数的候选值,并对所有可能的参数组合进行穷举搜索来寻找最佳的参数组合,而随机搜索则在给定的参数范围内随机采样一组参数,并评估其性能,以找到最优解。

(5) 自动机器学习(AutoML): 自动机器学习旨在自动化建模过程,包括特征选择、算法选择和超参数调优等。自动机器学习工具可以通过搜索不同的模型架构和参数组合来评估其性能,从而帮助策略调优和参数搜索。

(6) 多臂赌博机(Multi-armed Bandit): 多臂赌博机是一种用于在线优化策略的算法。在多臂赌博机问题中,代理需要在多个选择(臂)之间进行权衡,以最大化总体收益。该算法通常包括在探索(Exploration)和利用(Exploitation)之间进行平衡,以在不断试验新选择的同时,优先选择已知更好的选择。多臂赌博机算法常用于在线广告投放、推荐系统等领域。

综合运用上述统计或 AI 技术,可以使策略调优和参数选择更高效、更智能化,但仍需人工判断结果的合理性。

3.3.6 业绩归因分析

业绩归因分析可以运用 AI 技术进行增强:

(1) 利用自然语言处理技术分析业绩报告、管理层讨论等非结构化文本,识别影响业绩的关键因素。

(2) 应用机器学习算法建模分析各种定量指标对业绩的贡献度和相关性。

(3) 使用知识图谱技术分析业绩影响因素之间的关系,进行推理,以此来评估各因素的贡献。

(4) 利用强化学习等技术分析管理决策与业绩结果之间的关系,评估决策的价值。

(5) 应用文本挖掘技术处理社交平台、新闻等外部文本数据,识别市场反馈与业绩的相关性。

(6) 应用数据可视化技术,辅助交互式分析业绩影响因素。

(7) 利用仿真技术建立内外部环境模型,模拟评估业绩贡献情况。

(8) AI 技术可以从多个维度辅助业绩归因分析,结合人工判断之后,可以使流程和结果更全面、更准确、更智能化。

3.4 AI 驱动下的知识库搭建

量化策略的来源有很多,高质量的论文是量化策略的重要灵感来源之一。一般来讲,高质量的量化投资论文可以从谷歌学术、arXiv 等平台获取。不过获取论文仅仅是第 1 步,如何在获取论文之后,对论文进行快速阅读、分类、整理,并在之后的过程中快速检索并重新理解和阅读才是具有挑战性的环节。

目前,在 AI 的帮助下,量化研究者已经可以通过一些智能化工具高效地获取论文信息

或构建量化投资知识库。

开源项目 ChatPaper,能基于 arXiv 下载最新论文,并使用 ChatGPT 等预训练语言模型对论文进行提炼总结,生成固定式的摘要,辅助读者快速抓住论文要点。

自建企业级的中英文预训练语言模型,如 LangChain + ChatGLM。LangChain 支持基于行业 Embedding 的模型训练,可以更好地理解金融领域知识,构建私有化的知识图谱。

使用集成了论文理解功能的量化平台,如 BigQuant 的 QuantChat 组件。通过设计调整大语言模型的 Prompt,可以实现从论文阅读到策略回测的一体化。

市面上有不少量化平台构建了量化知识库供量化研究者使用,当然研究者也可以选择在本本地建立专属于自己的量化知识库。

本节将着重介绍如何着手搭建自己的量化知识库并使用。

3.4.1 LangChain 简介

LangChain 是一个强大的框架,旨在帮助开发人员使用语言模型构建端到端的应用程序。它提供了一套工具、组件和接口,可简化创建由大型语言模型 LLM 和聊天模型提供支持的应用程序的过程。LangChain 可以轻松管理与语言模型的交互,将多个组件链接在一起,并集成额外的资源,例如 API 和数据库。它不仅能实现通过 API 调用语言模型,还具备以下特性。

- (1) 数据感知:将语言模型连接到其他数据源。
- (2) 具有代理性质:允许语言模型与其环境交互。

LangChain 有很多核心概念,包括组件和链。组件是模块化的构建块,可以组合起来创建强大的应用程序。链是组合在一起以完成特定任务的一系列组件(或其他链)。

- (1) 模型(Models):LangChain 支持的各种模型类型和模型集成。
- (2) 提示(Prompts):包括提示管理、提示优化和提示序列化。
- (3) 内存(Memory):内存是在链/代理调用之间保持状态的概念。LangChain 提供了一个标准的内存接口、一组内存实现及使用内存的链/代理示例。
- (4) 索引(Indexes):与自己的文本数据结合使用时,语言模型往往更加强大——此模块涵盖了执行此操作的最佳实践。
- (5) 链(Chains):链不仅是单个 LLM 调用,还包括一系列调用(无论是调用 LLM 还是不同的实用工具)。LangChain 提供了一种标准的链接口、许多与其他工具的集成。LangChain 提供了用于常见应用程序的端到端的链调用。
- (6) 代理(Agents):代理涉及 LLM 做出行动决策、执行该行动、查看一个观察结果,并重复该过程直到完成。LangChain 提供了一个标准的代理接口,一系列可供选择的代理,以及端到端代理的示例。

具体可参考 LangChain 中文网提供的中文文档,LangChain 还提供了许多其他功能,包括提示模板和值、示例选择器、输出解析器、索引和检索器、聊天消息历史记录、代理和工具包等。

等。笔者的需求是从 2003 年到 2023 年金融量化论文，所以需要设置谷歌学术爬取的时间范围，可以按照时间年份逐年爬取并存储。搜索结果如图 3-7 所示。

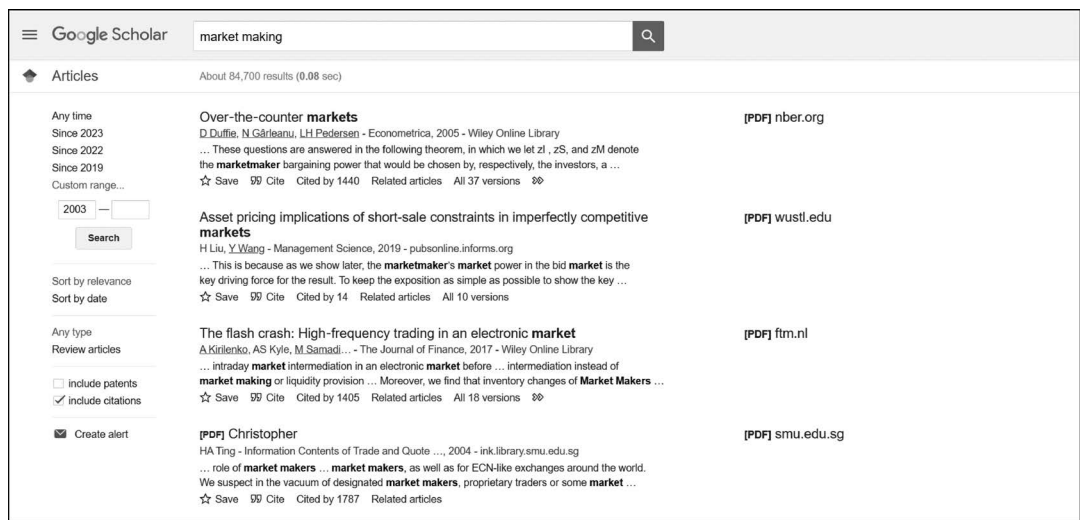


图 3-7 搜索结果(图片来源于谷歌学术)

- (2) 确定爬取范围：确定需要爬取的时间范围、关键词或其他过滤条件，以便限定爬取的论文范围。
- (3) 设置爬虫：使用适当的爬虫框架或工具(如 Python 中的 Scrapy 或 BeautifulSoup)编写爬虫程序，配置相关参数和规则。
- (4) 连接谷歌学术：编写代码以建立与谷歌学术的连接，并发送相应的搜索请求。
- (5) 解析搜索结果：解析谷歌学术返回的搜索结果页面，提取每篇论文的标题、摘要、作者等信息。
- (6) 下载论文内容：从搜索结果中获取论文的 URL 或其他标识，然后访问对应的网页或 PDF 文件，下载论文的全文内容。
- (7) 文本预处理：对于每篇论文的标题、摘要和全文进行文本预处理，如去除停用词、分词、词干化等。
- (8) 文本向量化：使用适当的文本向量化技术(如 Word2Vec、TF-IDF、BERT 等)将论文的文本内容转换为向量表示。
- (9) 存储到向量数据库：选择合适的向量数据库，将论文的向量表示和其他元数据存储到数据库中。
- (10) 建立索引：对存储在向量数据库中的论文数据建立适当的索引，以支持快速的相似度搜索和查询操作。
- (11) 查询和检索：使用数据库提供的查询功能，可以基于相似度计算或其他条件进行论文的查询和检索操作。

3.4.3 寻找高质量论文并下载

本节主要由两部分组成,分别是通过传统方法寻找高质量论文和借用 AI 工具寻找一簇高水平论文。

首先介绍如何用传统方法寻找高质量论文,在金融量化领域中寻找高质量论文是一个具有挑战性的任务,因为这需要综合考虑多个因素。以下是一些方法和指南,可以帮助读者在寻找高质量论文和顶级作者方面做出更明智的决策。

(1) 学术期刊和会议:寻找并关注金融量化领域内权威的学术期刊和会议,这些论文通常经过同行评审,确保研究质量较高。

(2) 引用次数和影响因子:考虑通过查看论文的引用次数和期刊的影响因子来评估其影响力和重要性。高引用次数和较高的影响因子可能是论文和作者质量的指标之一。

(3) 学术机构和研究中心:关注在金融量化领域内有声望和专业性的学术机构和研究中心。这些机构通常有顶级研究人员和专家,他们的论文可能是高质量的。

(4) 学术评价指标:了解和使用学术评价指标,如 H 指数、G 指数、引用报告和学术排名。这些指标可以提供关于作者和论文影响力的参考。

(5) 领域专家意见:寻求金融量化领域内的专家和学者的建议和意见,他们对该领域的发展和前沿研究有深入了解。

(6) 学术网络和合作:研究学者的学术网络和合作关系,通过查看他们与其他知名研究人员的合作情况,可以获得关于其研究声誉和质量的线索。

(7) 综合评估:综合考虑上述因素,并结合自身的研究兴趣和需求,进行评估和判断。高质量论文和顶级作者的定义因人而异,所以重要的是根据自己的标准和目标进行判断。

国外有一些网站针对各种文献之间相互引用和影响因子及论文研究领域的聚合研发了一些人工智能的工具,可以帮我们快速找到需要的论文。此处介绍两个工具:

第 1 个工具是网站类工具,可以生产论文链接图谱,如图 3-8 所示,方便我们快速找到中心枢纽的论文及影响因子比较大的论文。该图谱可以非常方便地让我们查看论文之间相互引用的关系及该论文的权重。单击每个节点可以查看详细的引用数、收录机构及下载论文的链接等。

第 2 个工具是 Research Rabbit,如图 3-9 和图 3-10 所示。

3.4.4 利用 ChatGPT 进行批量论文粗读

笔者按照 GitHub 上的托管代码 Chatpaper 在云端服务器自建了一个 LLM 进行论文粗读,也就是进行论文筛选,该模式采用了一个模板结构针对一篇论文进行总结。以下是一篇论文的总结样例。

Summary_Result:

(1) Title: Optimal high frequency trading with limit and market orders (通过限价单和市价单优化高频交易)。

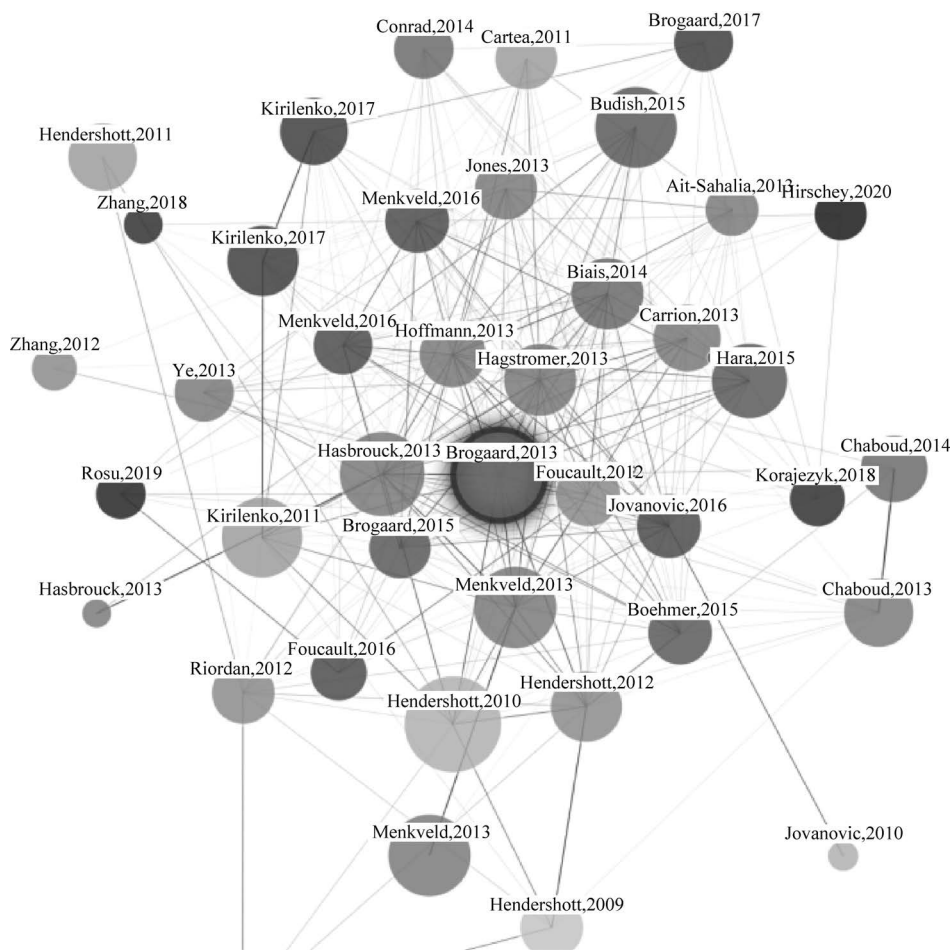


图 3-8 论文链接图谱(图片来源于 connected papers 网站)

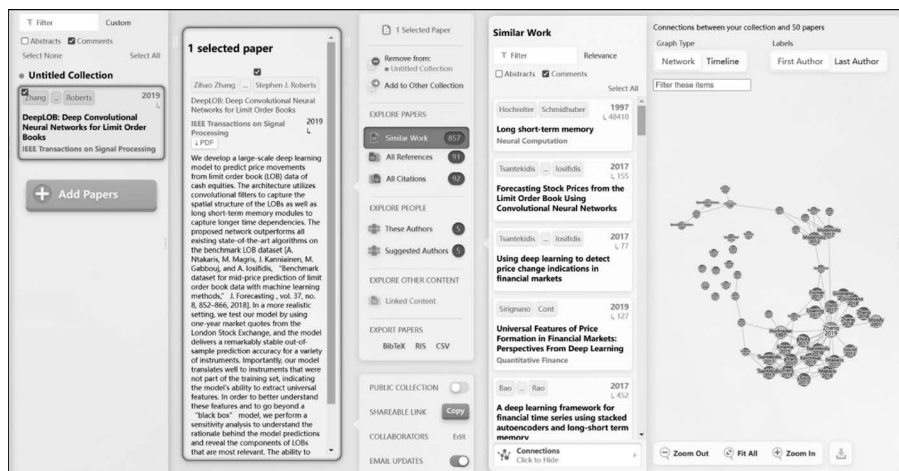


图 3-9 Research Rabbit 界面(图片来源于网络)

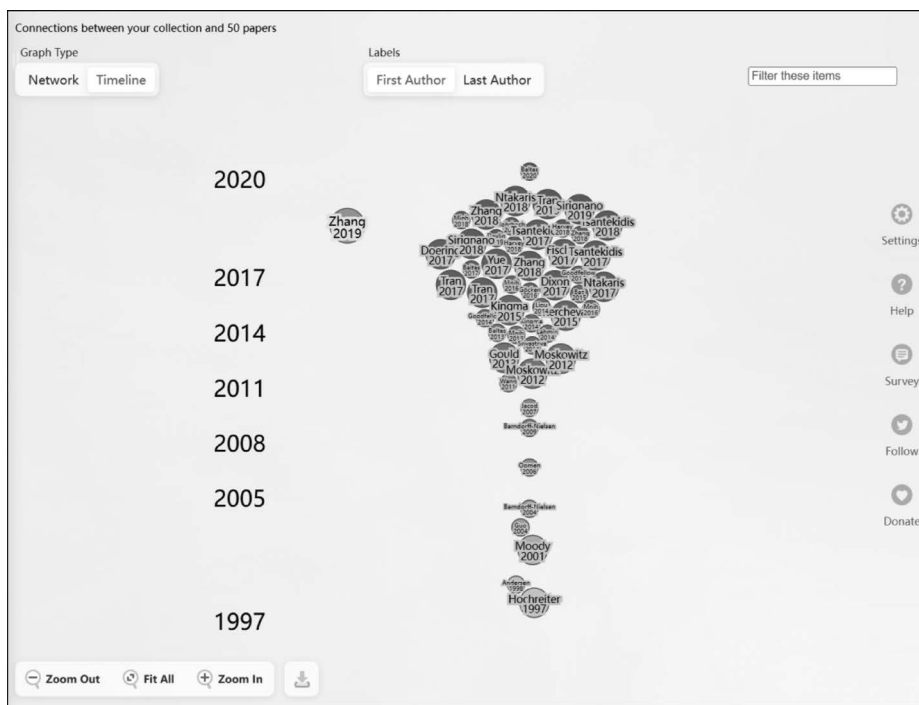


图 3-10 Research Rabbit 论文图谱(图片来源于网络)

(2) Authors: Fabien Guilbaud, Huyên Pham.

(3) Affiliation: EXQIM and Laboratoire de Probabilités et Modèles Aléatoires CNRS, UMR 7599 Université Paris 7 Diderot(法比安·吉尔鲍德 and Huyên Pham, 巴黎第七大学概率和随机模型实验室)。

(4) Keywords: Market making(市场做市商), limit order book(限价订单簿), inventory risk(库存风险), point process(点过程), stochastic control(随机控制)。

(5) URL: Paper-<https://arxiv.org/abs/1106.5040> GitHub-None。

Summary:

(1) 本文研究了限价和市价订单下的最优市场做市商(Market making)策略。作者主要研究了市场做市商交易时面临的限价订单执行风险和库存风险对利润的影响。

(2) 过去的方法主要集中在执行市价交易,相较之下本文提出了一种新的解决方案,该方案可以在交易量较大时有效地利用限价订单提高执行优先权并减轻市场波动带来的风险。该方法的提出有足够的动机和理论支持。

(3) 本文提出了一个框架来研究市场做市商的参与策略。该框架以 Markov Chain 作为限价订单簿中买卖价差的模型,由程序员指定每个时刻可下订单的时间单位,从而得出该单位下一步限价订单簿的状态和各种激活该状态的时间分布。作者定义了一个包括控制策略和交易动机的交易模型,并针对该模型提出了一个动态规划算法,最终构建了混合策略的

最优控制问题。

(4) 本文建立的交易模型可以产生最小的执行和持仓风险来获得最大的收益,该模型已经在真实数据上得到了广泛应用,但尽管方法具有一定的应用价值,但作者并未给出该方法相对于以前的方法的实验结果。

```
prompt_token_used: 1516 completion_token_used: 674 total_token_used: 2190
response_time: 20.267 s
conclusion_result:
```

Conclusion:

(1) 本文对于高频交易中市场做市商的最优策略进行了研究,提出了一个包括控制策略和交易动机的交易模型,并针对该模型提出了一个动态规划算法,最终构建了混合策略的最优控制问题。该方法的提出对于高频交易领域的研究和实践具有一定的意义。

(2) 创新点: 本文提出了一种在限价和市价订单下实现最优市场做市商策略的模型和算法,相较于以前的研究,该方法能够更好地应对交易量增大、波动加剧时的风险。性能: 本文的模型和算法在真实数据上得到了广泛应用,能够产生最小的执行和持仓风险来获得最大的收益。

(3) 工作量: 本文的工作量较大,涉及较为复杂的数学模型和算法,但是对于高频交易领域的研究和实践来讲具有一定的推动作用。

```
prompt_token_used: 971 completion_token_used: 344 total_token_used: 1315
response_time: 10.454 s
```

3.4.5 利用 ChatPDF 进行论文精读

ChatPDF 是一个人工智能工具,它可以帮助用户快速阅读 PDF 文件。用户只需将 PDF 文件上传到 ChatPDF,然后就可以开始向它提问有关该 PDF 文件的任何问题。

ChatPDF 非常适合快速提取各种论文的摘要,并且支持中文输出。这样,用户就可以快速地了解论文的主要内容,从而提高阅读论文的效率。可以访问 ChatPDF 的官网了解更多信息。

上传一个 PDF 文件,该工具直接给出 3 个 Example Questions,单击小图标,该工具会直接提供答案并给出该答案引用的页码,如图 3-11 所示。



图 3-11 ChatPDF 界面(图片来源于网络)

当然也有很多类似的工具,例如 ChatDoc 也提供了文件和人进行交互的方法。

3.4.6 建立私有知识库并进行交互

根据 GitHub 上开源项目建立的本地化的知识库,项目网址可从目录处二维码扫描获取。该项目可以非常方便地进行私有知识库的交互,该项目 Embedding 默认选用的是 GanymedeNil/text2vec-large-chinese,LLM 默认选用的是 ChatGLM-6B。依托上述模型,本项目可实现全部使用开源模型离线私有部署。过程包括加载文件→读取文本→文本分割→文本向量化→问句向量化→在文本向量中匹配出与问句向量最相似的 Top k 个匹配出的文本作为上下文和问题一起添加到 prompt 中→提交给 LLM 生成回答。流程如图 3-12 所示。

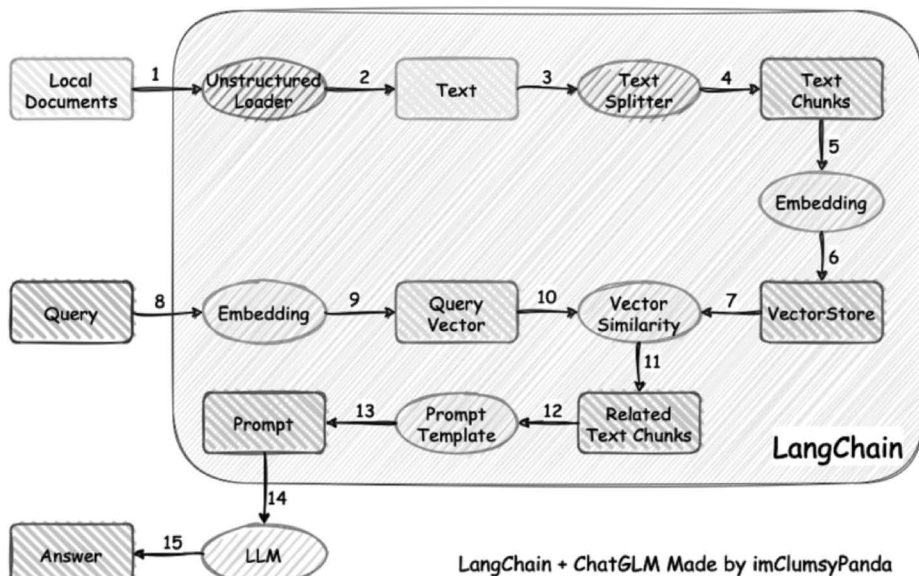


图 3-12 LangChain 处理流程 (图片来源于网络)

从文档处理角度来看,实现的流程如图 3-13 所示。

按照 GitHub 项目提示的部署好项目后,可以将需要解读的 PDF 文档上传到一个项目的子目录中,然后就可以针对一系列的文档进行交互,快速提取文档集合中的知识。实现以上的操作可以在 Webui 的界面下实现。

以下是该项目的部分操作指南:

运行前自动读取 Configs/model_config.py 文件中 LLM 及 Embedding 模型枚举及默认模型设置运行模型,如需重新加载模型,则可在重新选择“模型配置”后单击“重新加载模型”按钮进行模型加载。

可手动调节保留对话历史长度、匹配知识库文段数量,可根据显存大小自行调节。

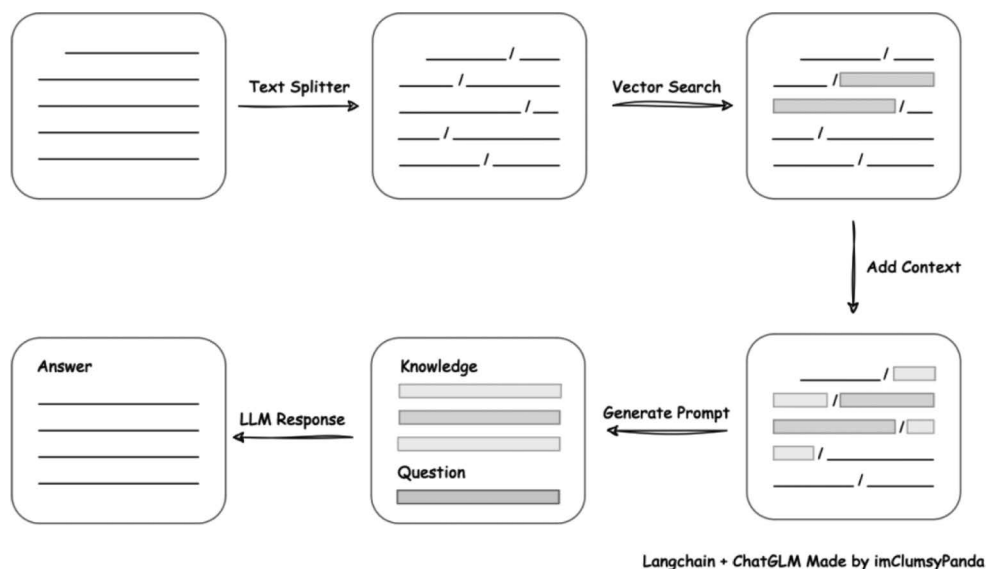


图 3-13 LangChain 回答流程 (图片来源于网络)

“对话”具备模式选择功能,可选择“ LLM 对话”与知识库问答模式进行对话,支持流式对话。

添加配置知识库功能,支持选择已有知识库或新建知识库,并可向知识库中新增上传文件/文件夹,使用文件上传组件选择好文件后单击“上传文件并加载知识库”按钮会将所选上传文档数据加载至知识库中,并基于更新后的知识库进行问答。

知识库测试 Beta,可用于测试不同文本切分方法与检索相关度阈值设置,暂不支持将测试参数作为“对话”设置参数。