开启证券资讯 3.0 时代——证券知识图谱

白雪，熊昊

上海证券交易所　资本市场研究所

E-mail: xbai@sse.com.cn

摘要：互联网+时代的到来标志着互联网从一个工具变成了一个基础性的设施，在互联网+时代，万物通过互联网进行互联，互联网的基础性地位日显重要，已经渗透到包括金融、物流、电子商务、工业生产等各个领域。互联网以信息作为其载体及表现形式的特性，与金融行业具有天然的融合性。金融证券行业对信息的分析与处理方法的探索从来没有停止过。从早期报业时代的人力分析，发展到了目前以信息技术手段进行海量金融数据分析，其分析方法、分析手段发生了翻天覆地的变化。然而上述对信息的分析方法仍然存在着缺陷。大数据处理中的数据孤岛问题、作用链传递的问题仍然突出。而证券资讯 3.0 时代正是为了解决这一问题应运而生。证券时代资讯 3.0 到来的表现形式为——证券知识图谱在证券行业的应用。2012 年，Google 正式发布了“知识图谱”（Knowledge Graph）——可以将搜索结果进行知识系统化，任何一个关键词出发都能获得完整的知识体系。建立证券行业的知识图谱通过对海量证券结构与非结构化数据进行知识抽取与分析，建立证券行业专业知识互连图，为证券行业智能分析、程序化交易、市场监管等活动打下了坚实的基础。

关键字：互联网+；知识图谱；证券资讯 3.0。

1 引言

互联网+时代的到来标志着互联网从一个工具变成了一个基础性的设施，在互联网+时代，万物通过互联网进行互联，互联网的基础性地位日显重要，已经渗透到包括金融、物流、电子商务、工业生产等各个领域。互联网以信息作为其载体及表现形式的特性，与金融行业具有天然的融合性。金融行业从本质上而言，就是由不同的数字与信息去表达金融资源的时间与空间特性，通过对其信息进行处理，完成不同金融资源的时间及空间的匹配，以达到资源效用最大化的目的。而在不同金融资源的时间及空间匹配过程中，以信息分析和信息处理为核心的风险控制手段起到了至关重要的作用。

互联网+金融将对金融行业产生颠覆性的影响，其表现形式为互联网金融，但实质上而言是信息流动与分析的手段重塑了金融行业的风险控制模式。任何金融的行为都可以看成是一个风险控制的行为，其抽象过程是一个尽量降低风险、提高收益的过程，而要实现这个过程，则必须充分合理的分析金融资源后面所蕴含的各类信息。

金融证券行业对信息的分析与处理方法的探索从来没有停止过。以股票投资市场为例，早期受到分析手段及资讯传导速度的限制，人们以分析结构化数据，例如股票的成交量、成交价格为主；在公司的基本面方面，则以分析公司的财务结构数据为主。在报业时代，受信息更新速度、传播速度的影响，通过对非结构的文本数据包括并不多；与此同时，报业时代产生的数据量并不大，由人工分析足以满足业务应用需求。在信息时代，一方面随着互联网时代的到来，资讯的生产方由专业媒体变成了大众，各类关于公司、市场的信息由不同的人士生成并发布，数据量空前丰富；另一方面，空前丰富的数据体量使得人工分析变得越来越不现实，则信息技术的成熟、应用成本的降低使得将信息技术应用于金融非结构化数据的分析服务成为可能。 在这个时期，证券行业纷纷通过搭建各类分析平台来对结构与非结构化数据进行采集与分析。

然而上述对信息的分析方法仍然存在着缺陷。首先，目前互联网已经进入到了互联网+时代，万物互联已经成为时代的主流，而上述的信息分析方法将一个个信息点进行孤立的分析，形成一个个信息分析孤岛，其表现形式为对单一问题、单一信息分析较为全面，但对多个问题、多个信息的关联分析等能力较为欠缺，分析结果零散；其次，表现为查询结果不够智能，只能就查询者的某个问题回答相应的答案，而不能够就问题所描述的知识结构完整全面的展示给查询方，互联网+时代万物互联的理念不相适应。对海量结构与非结构化金融大数据的智能分析，将金融数据所展示的数据以万物互联的理念进行处理与展现，催生了证券资讯业 3.0 时代的到来。

证券时代 3.0 到来的表现形式为 GOOGLE 公司推出的知识图谱在证券行业的应用。为了让用户能够更快更简单的发现新的信息和知识，给用户提供有完整知识体系的搜索结果，2012 年，Google 正式发布了“知识图谱”（Knowledge Graph）——可以将搜索结果进行知识系统化，任何一个关键词都能获得完整的知识体系。在处理一词多义这样的问题时，Google 知识搜索重新使用了在上世纪五六十年代首次提出的的语义网络的想法，

那是对人类意识在大脑中可能的编码信息做出的最早猜测。取代词与词之间简单的关联，采取信息编码上的唯一的对应实体。当所有的地点，人物和关系相互关联，这些网络便开始像一个巨大的蜘蛛网。从本质上讲，Google 正试图重塑互联网，提供一个更加智能的信息获取渠道。

在将智能大数据处理、知识图谱的应用方面，美国以 Kensho 公司为代表，而中国则以上海证券交易所的知识图谱项目为典型案例。2014 年 11 月 27 日，高盛向金融数据服务商 Kensho 投资 1500 万美元以支持该公司目前正在进行的数据平台分析开发计划。Kensho 目前正在研发一种针对专业投资者的大规模数据处理分析平台，该平台可以快速、大量的进行各种数据处理分析工作并且能够实时的回答投资者所提出的复杂的金融问题。Kensho 将撼动金融分析行业，他们想要将金融市场的一些专业知识交到大众手中，而此前仅仅只有一些顶尖级的对冲基金和银行使用这些专业知识来利用短期的市场无效赚取大量利润。Kensho 在做的工具将是使每位金融专业人士，不仅是定量分析师、程序员和数据科学家来提问和解答全球事件的难度问题和它们对于证券价格的影响。例如，当三级飓风袭击福罗里达州时，哪支水泥股的涨幅会最大? 当油价高于 100 美元一桶时，中东政局动荡会对能源公司的股价产生怎样的影响？要回答此类问题，对冲基金的分析师队伍要花上数天的时间，前提条件是他们可以找到所有这些数据。但 Kensho 可以通过扫描药物审批、经济报告、货币政策变更、政治事件、以及这些事件对地球上几乎所有金融资产的影响等 9 万余份资料，立马就为 6,500 万个问题找到答案，让技术给市场带来透明度。随着 Kensho 这类基于海量行业大数据扎根于行业的服务平台的出现，证券资讯行业正面临着一场深度变革。传统的证券资讯，是各种原始资讯的收集，清洗，整理，最后以分门别类的方式展示或推送给用户。这其中，消息与消息之间的关系，消息与股票之间的关系，人与人之间的联系，人与股票之间的联系等等，大多数难以呈现。这些数据和资料，仅仅简单的分类存于仓库中，而他们之间的关联情况，则需要从业人士自己发掘和建立。随着人工智能技术的发展，大数据时代各类数据资讯的充盈，这无疑为建立证券行业万物互联——即，构建证券行业知识图谱，提供了充沛原材料和坚实的工具。

而在知识图谱的应用方面，国内证券市场则以上海证券交易所设立的“基于知识图谱的沪港两地情绪指数关键技术研究”这一联合研究课题为开端。该课题是国内首批将知识图谱应用于证券金融领域的科研项目，旨在利用知识图谱这一先进的国际理念与技术，更好的服务于市场核心机构从事市场监管、市场参与机构进行盈利等。就具体而言，知识图谱在证券行业的应用包括以下几个部分。

对交易所等市场核心机构而言，在监管实践过程中面临着信息爆炸的问题。目前，上市公司的数目众多，基于互联网平台的股吧、论坛、门户网站、微信、微博等每时每刻也在产生着大量的信息，上述信息将极有可能会对股价产生影响。就交易所而言，首先希望通过知识图谱建立起上市公司之间的关联关系，立图建立一个全谱系的上市公司关联图。其次，希望能够对上述信息平台的数据进行实时的挖掘与采集。而在具体业务应用方面，当监控到市场价格出现波动时，可以就股价出现异动的股票在知识图谱中追溯其异动产生的根源，尽早的采取各种监管措施，维护投资者的合法权益。同时，知识图谱可以自动的学习实体之间的隐含关联关系，这对于打击内幕交易等监管活动会起到重要的帮助。

对于以证券公司、公私募投资机构为主体的市场参与者而言，与传统环境相比，在互联网时代，其在进行投资决策过程中面对的信息量更大，对信息的快速分析与理解将有助于更好的在瞬息万变的市场中获利。通过上市公司领域知识图谱的建立，证券公司可以进行快速的事件套利。具体而言，通过对市场即时事件信息进行实时的导入，通过证券行业知识图谱将所有重点相关联的行业、版块、公司、股票以及个人进行影响值排序，对上述信息可能产生的正面或负面影响进行实时的分析并得出相应的结论，使得机构可以先于市场其他参与者发掘出潜在关联方并全面的分析出事件波及影响层面，从而快速作出投资决策实现盈利或止损。因此，将基于知识图谱的信息分析服务产品作为一个增值服务提供给证券公司的客户，也可以在互联网证券时代为证券公司更好的理解用户需求，提供更为个性化更为智能化的信息服务，以大大增强客户粘性。

图 1 简化地展示了证券行业知识图谱的部分关联情况。具体来说，证券行业需要研究的标的覆盖了各行各业，行业与行业之间可能有父子关系和上下游关系；产品可以属于某些行业，产品层的实体可以与行业层的实体建立关联，同时产品与产品之间可能存在原材料等关系；公司主营生产若干产品，可将公司与其对应的产品之间建立关联，同时，公司与公司之间也有母公司-子公司、参股、战略合作等关系；公司的股东、法人等也是一个重要的需要监测的实体对象，某些股东之间可能存在亲缘、密友、同学等关联关系。

总之，证券行业是个万物互联的行业，只有逐步建立起千丝万缕的联系，并将各种联系的强度进行量化，才能大大提高行业信息利用的精准度和可信度，迅速跟踪行业事件引发的各种蝴蝶效应，从而拓宽信息掌握的广度、精度与速度，强化信息组织和处理能力，使得大数据确确实实能为行业决策提供坚实的数据支持和分析依据，开启证券行业智慧资讯 3.0 时代。

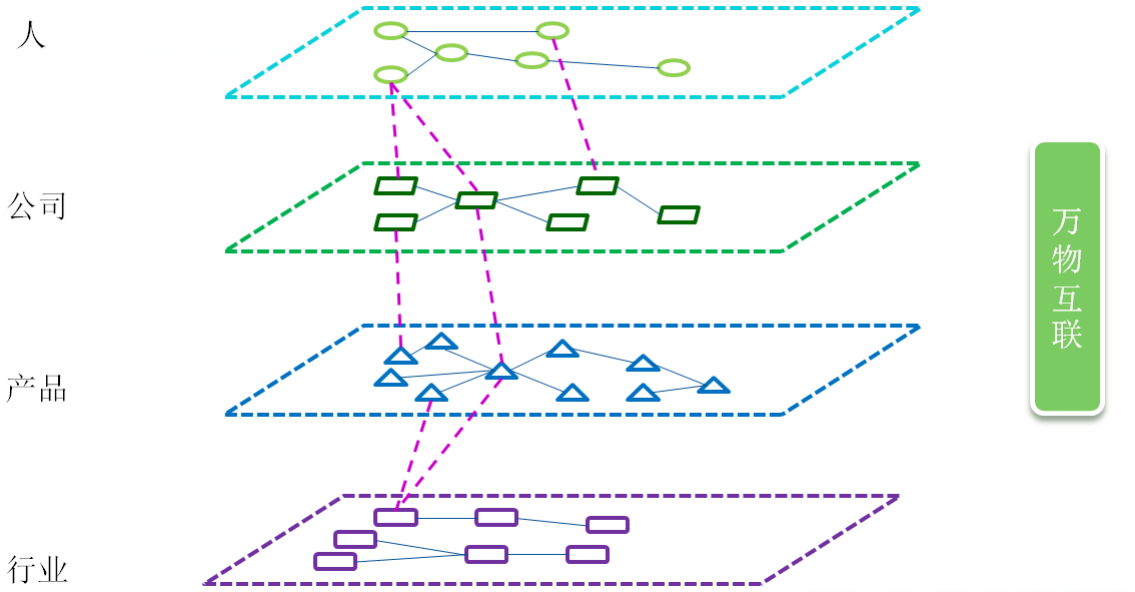

图 1 证券行业知识图谱

2 知识图谱概述

为了让用户能够更快更简单的发现新的信息和知识，给用户提供有完整知识体系的搜索结果，2012 年，Google 正式发布了“知识图谱”（Knowledge Graph）——可以将搜索结果进行知识系统化，任何一个关键词都能获得完整的知识体系。在处理一词多义这样的问题时，Google 知识搜索重新使用了在上世纪五六十年代首次提出的的语义网络的想法，那是对人类意识在大脑中可能的编码信息做出的最早猜测。取代词与词之间简单的关联，采取信息编码上的唯一的对应实体。当所有的地点，人物和关系相互关联，这些网络便开始像一个巨大的蜘蛛网。从本质上讲，Google 正试图重塑互联网，提供一个更加智能的信息获取渠道。

知识图谱用于广义搜索领域有以下几点优势：(1) 找到正确的结果。由于一个关键词可能代表多重含义，所以知识图谱会将最全面的信息展现出来，让用户找到自己最想要的那种含义。(2) 最好的总结。有了知识图谱，搜索引擎可以更好的理解用户搜索的信息，并总结处相关的内容和主题。(3) 更深、更广。由于“知识图谱”会给出搜索结果的完整知识体系，所以用户往往会发现很多不知道的知识。

什么是知识图谱？知识图谱是一种大规模的知识表示形态，本质上是一种语义网络。知识图谱可被看作是一张巨大的图，节点表示实体或概念，边则由属性或关系构成。知识图谱旨在描述真实世界中存在的各种实体或概念及其关系，一般用三元组表示，主要研究知识的表示，实体的关联，概念实例化等。

互联网上可用于建立知识图谱的数据源包括百度百科、维基百科等百科类网站，专业方面的网站，以及搜索日志等，中文百科类的站点，如百度百科等，的结构化程度远不如维基百科，能通过信息框获得 AVP 的实体非常稀少，大量属性-值对隐含在一些列表或表格中，需要针对性的做处理和提取。搜索日志也可以作为知识图谱构建的重要资源。一条搜索日志形如<查询，点击的页面链接，时间戳>。通过挖掘搜索日志，我们往往可以发现最新出现的各种实体及其属性，从而保证知识图谱的实时性。这里侧重于从查询的关键词短语和点击的页面所对应的标题中抽取实体及其属性。选择查询作为抽取目标的意义在于其反映了用户最新最广泛的需求，从中能挖掘出用户感兴趣的实体以及实体对应的属性。而选择页面的标题作为抽取目标的意义在于标题往往是对整个页面的摘要，包含最重要的信息。

现有的大规模知识图谱包括：

(1)Yago：1 千万实体，35 万类别，1.8 亿事实，100 种属性，100 语言。

(2)Dbpedia：4 千万实体，250 类别，5 亿事实，6000 种属性。

(3)Freebase：2 千 5 百万实体，2000 主题，1 亿事实，4000 种属性。

(4)谷歌知识图谱：5 亿实体名字，35 亿条事实。

(5)NELL：3 百万实体名字，300 类别，500 属性，100 万事实，1 千 5 百万学习规则。

以上知识图谱主要以英文搭建，而现有的中文知识库，以面向自然语言分析和面向行业业务需求为区分，可分为如下两类：

(1)传统的语言类知识库，基于人工编写方式，构建了一系列的中小规模中文知识库:

- 知网(HowNet)[1]
- 《同义词词林》[2]
- 概念层次网络(HNC)[3]

(2)大规模事实类知识库，针对自身业务需要建立

- 百度知心，优化搜索
- 搜狗知立方，优化搜索[4]
- 阿里巴巴知识库(商品知识库)

然而这些都不能直接构成知识图谱，仅仅是从各种类型的数据源抽取构建知识图谱所需的各种候选实体及其属性关联，形成了一个个孤立的抽取图谱（Extraction Graphs）。为了形成一个真正的知识图谱，我们需要将这些信息孤岛集成在一起。

知识图谱的维护和更新可以采用众包反馈机制：除了搜索引擎公司内部的专业团队对构建的知识图谱进行审核和维护，它们还依赖用户来帮助改善图谱。具体来说，用户可以对搜索结果中展现的知识卡片所列出的实体相关的事实进行纠错。当很多用户都指出某个错误时，搜索引擎将采纳并修正。这种利用群体智慧的协同式知识编辑是对专业团队集中式管理的互补。

3 系统架构与主要模块

构建证券领域的知识图谱系统，需要搭建网络爬虫、文本分析、社交网络分析、证券知识图谱构建和情感分析四个模块，图 2 是整体架构图，接下来将分别进行介绍各个模块。

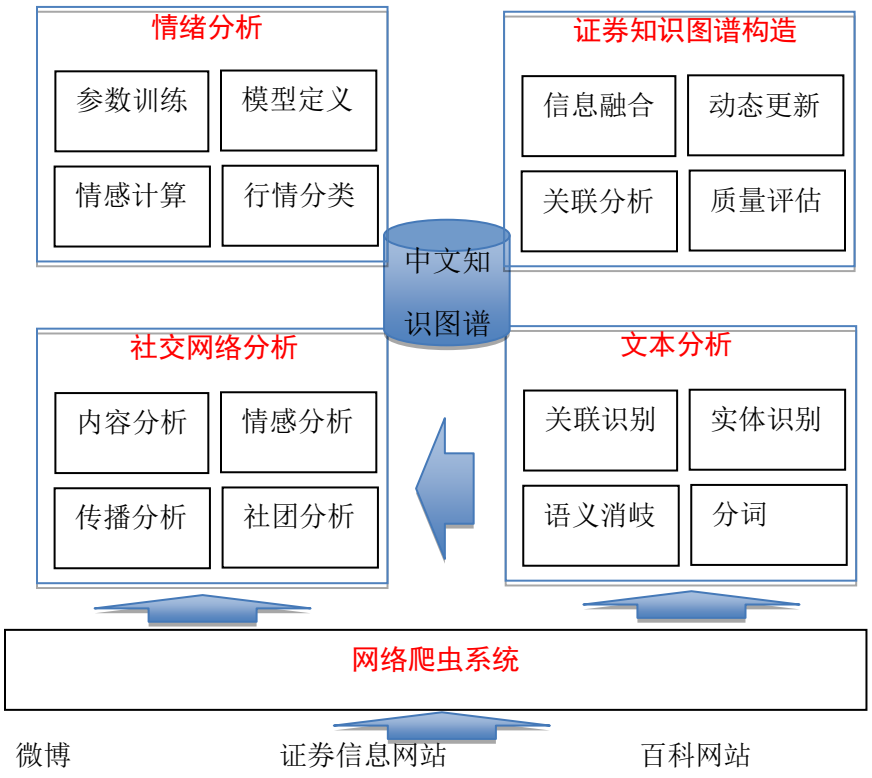

图 2 系统架构

3.1 数据源

证券领域知识图谱的主要数据来源来自于如下系统：

百科类网站：百度百科、WIKI 等。主要获取证券领域的主要概念，并分析概念之间关联关系。

证券信息网站：证券交易所、期货交易所、彭博等，主要获取交易行情数据、证券产品的说明文档、公告等。

□ 微博等社交网站：主要获得民众、股民对相关金融产品的评价和舆论。

3.2 网络爬虫系统

系统将开发一种自动检测屏蔽的通用分布式爬虫系统，该系统可运行在 100 台机器的集群上。该系统基于 MapReduce 分布式计算框架，目标是最大程度地利用集群的网络资源和计算能力，快速地管理并进行对大批量 Web 网页的抓取，同时有强大的扩展性以方便地支持各种特殊的抓取任务，核心技术是在不需要人工介入的情况下自动检测是否因大量访问而被服务器封锁，其任务调度算法能依此规避封锁而达到稳定快速的抓取效果。

该系统的特性有：

多任务：系统可以同时管理多个抓取任务，并且能依据任务的优先级对每个任务进行智能的切分和调度。可以通过远程连接或直接控制系统核心节点以方便地添加或移除任务。

高效率：系统能最大程度地利用所有的机器资源，基本调度算法实现了自动的负载均衡，不会在某台爬虫机器上发生堆积。除非没有任务、机器故障或服务器封锁，否则每台爬虫机器都会在工作状态。

分布性：同一个任务会尽量分配给多个爬虫机器，以避免某台爬虫机器对某一个网站的高负载访问而被屏蔽。

容错性：系统可以在网络错误、机器故障，甚至因为服务器封锁而返回错误的页面等情况下仍保证每个任务成功完成，其错误检测和自动重试机制会完成这个条件。并且系统拥有快照机制可以在发生致命错误时恢复。系统拥有心跳检测机制以控制每台机器的状态。

可扩展性：分为任务类型的可扩展性和系统硬件的可扩展性。系统拥有着广泛性的接口以方便地支持各种各样的抓取任务，如对 ajax 动态网页的抓取、对视频站视频的抓取等比较特殊的任务，并且任务调度策略也可以进行扩展。系统可以在运行中任意增加或删除爬虫机器，而其上运行的任务不会因此受到影响。

智能屏蔽检测：系统通过对下载到的网页进行分析，运用异常状况检测算法，自动估计该任务当前是否被服务器屏蔽而获得了错误的结果。这个结果将被系统的调度算法分析以使该任务被分派到另外的爬虫机器上，以最大程度地减少因为服务器方原因和受到的影响。

图 3 是系统架构图。Master 中包括任务指派模块、任务调度模块和错误处理与故障恢复模块。每个抓取任务被称为一个 Job，每个 Job 有一个与其相关联的爬虫程序，如对某个 Ajax 动态网页的抓取任务就需要一个特制的 Ajax 爬虫。Master 上保存有多个爬虫，并且可以方便地往里面添加新的爬虫类型，以此来支持各种各样的抓取 Job 的类型。任务调度模块中的任务池保存当前系统中所有正在执行的 Job 及其状态。被分发到 Slave 的任务是用这个池中获取的。

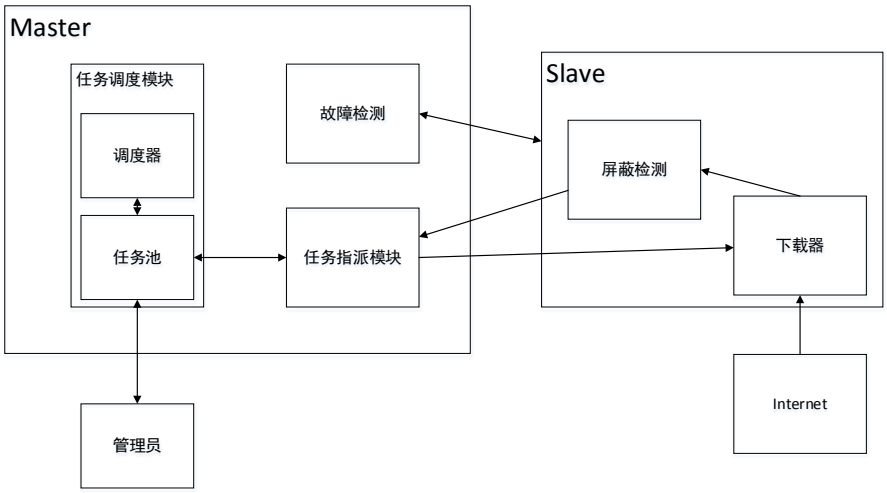

图 3 网络爬虫系统框架

3.2.1 系统基本架构

系统采用星型网络架构，即一台核心节点（以下称为 Master）控制集群中其他的节点（以下称为 Slave，

从机)。Master 负责管理整个集群，包括集群故障状态和被封锁状态的监控、任务管理和调度分发。Slave 负责具体的抓取工作，即访问网络和下载页面的过程。Slave 只作为执行者，只和 Master 进行通信，并且不保存有关任务的信息。所以 Slave 的添加、删除、故障都不会影响系统的正常运行。Master 通过快照机制来处理单点故障（即 Master 故障导致集群瘫痪的情况)。

3.2.2 任务指派机制

Master 通过心跳机制检测每一台 Slave 的状态，当它发现有 Slave 处于空闲状态，它会从 Job Pool 中取出一个 Job，将其分割成若干平行小任务，并把这些小任务按照分配策略分配给空闲的工作机以确保最大程度利用计算资源。这里每个任务称为该 Job 的一个 Task。每台 Slave 在完成其任务后会向 Master 报告任务的完成状况，Master 记录每个 Task 的状况，可能是待分配、正在执行中、已完成、或者是已出错，对于已出错的任务，Master 会将其分配给另一台 Slave 执行。最终保证整个 Job 能成功的执行。

某些抓取任务可能有 Reduce 流程，如可能需要跟踪获得的所有页面中的超链接，此时 Master 需要获得这些超链接用作下个 Job。对于这些任务，Master 会向 Slave 收取这些结果，收集完成之后 Master 可以根据这些结果执行 Reduce 任务。

3.2.3 任务调度机制

对 Job Pool 中的每个 Job，Master 除了保存它们由用户输入的配置信息外，还保存有它们的域名、被封锁后的冷却时间、执行状态、以及优先级，调度模块负责在当有空闲 Slave 时，对 Job 进行恰当合理的调度，以达到以下几个要求：优先度高的比优先级度低的有更多的抓取资源，即有更高的机会被指派开始抓取。一个 Job 会被尽量平均地分配到各个 Slave 上，以尽量避免某台 Slave 因为高速连续访问而被服务器封锁。若发现某个服务器（某个域名）封锁了某个 Slave，那么在一段时间内绝对不会将和该服务器（域名）有关的 Job 指派到这个 Slave 上。系统在发现空闲 Slave 后，先寻找 Job Pool 中优先级最高的 Job，并尝试将该 Job 的一个不是完成状态的 Task 指派给该 Slave（Recall: Task 是一个 Job 中的一个小部分)。这里的尝试指的是要评估是否能够将该 Job 的 Task 指派给 Slave。这个评估包括该 Job 的服务器屏蔽了该 Slave 的可能性、Slave 是否能够运行该 Job、Slave 是否正常运行等。若评估结果是该 Slave 可以运行此 Job，则此 Task 在该 Slave 上开始运行，否则系统尝试优先级次高的 Job。如此继续下去，或者该 Slave 被指派成功，或者没有任何适合该 Slave 运行的 Job，该 Slave 进入暂时休眠，一段时间后再次尝试。

3.2.4 错误处理与故障恢复机制

从上面的任务指派机制可以看出，只要某个 Job 的某个 Task 没有被标记为已完成，那么这个 Task 将会作为“以前的未完成 Task”被再次指派知道完成为止。这里没有被标记为已完成的原因可能是这是一个新的 Task，可能是 Slave 报告了一个错误或者屏蔽，也可能是 Slave 由于当机而没有返回报告任何结果。这个机制保证了这类故障不会影响系统的正常运行，每一个 Job 在完成时总能够保证其分割成的每一个 Task 都是完整地完成了的。

星型结构能保证对于非中心节点以外的节点故障都很容易处理，但是此结构存在单点故障，即一旦中心节点不幸故障，那么整个系统都将停止运行。为解决这个问题，该系统拥有快照机制，中心节点每隔一段时间就会保存一次 Job Pool 的状态。这样不论是 Master 当机还是断电等突发性的大故障，系统都能很容易地返回故障前最后一个快照的状态，这个状态通常只在几十分钟之前。所以系统能保证其健壮性，在绝大多数故障下系统都能保证其中的 Job 能成功完成。

3.2.5 全自动屏蔽检测

这个模块的目标是能及时发现当前抓取任务是否被禁(Ban)，如果发现那么调度系统可以进行调度以避免系统做无用功，也避免因为过度访问而使得其 IP 被服务器永久封锁。更重要的是这个机制应当是不需要人工介入的以节省大量人力成本。

我们将抓取到的网页分成 3 种类型，分别是正常页面、错误页面和 Ban 转向页面。注意到有些网站的 Ban 转向页面并非输入验证码，而是转到首页或某些特定页面，同时分析网页文本内容的方法代价也更大，而且其内容复杂性使得分析错误率相对较高（如某些网页中可能出现“验证码”等字样)。

系统通过一种基于检测网页特征的分布异常的算法来辨认被封锁页面。我们假定对同一个任务（同一个域名下的网页）Ban 页面和错误页面具有以下性质：Ban 页面总是相似（即它们的字符数相差不大，且其任 2 个的编辑距离很小）且连续的

(1) 错误页面总是相似且不连续的

(2) 对正常页面，其特征波动较大

(3) 多个实验结果可以支持这些假定。

并且我们发现 Ban 页面通常可能是输入验证码的页面或转到首页等默认页面，也可能会比较复杂，有比较大的 Size，只根据 Size 判断不可行。但是除了转回页面和输入验证码的标记不一样外，它们整体框架相似，

即它们的 Token 编辑距离（将 Web 页面按照 Html Tags 分割成 Token 列表之后的编辑距离，可以体现网页的框架格式信息而一定程度上忽略文字等具体信息）会很小。而连续性表现为一旦被 Ban，后面所有页面都会转到该页面。

3.3 文本分析系统

文本分析系统主要针对各种不同类型的文本数据进行分析和理解，以获取其中金融产品实体信息及其关联关系。主要研究内容包括：

(1) 分词

项目将建立一个面向证券行业的分词库，在该词库将结合现有的词库并增加证券领域专业词库的内容。并将该词库融合到分词系统中，从而提升证券领域文档的分词效果。

(2) 实体识别

该模块主要实现对文本文档中证券相关实体的识别。项目将结合证券知识图谱系统建立证券实体库，并建立证券名词的同义词表，从而从文本数据中抽取相关的实体。并结合指代词分析，获取金融实体信息在文本中出现的位置。

(3) 关联分析

该模块主要实现对文本文档中实体之间关系的抽取。本项目中将主要针对具有部分结构信息的文本和自由文本两类文档进行处理，两者的处理方式有很大的不同。

具有部分结构信息的文本。在百科类网页或交易所的公告文档中，实体之间是有一定结构的，项目将研究基于自动机描述的实体关联关系抽取技术。用户可以根据文档的结构定义实体间关系的提取模式，并开发面向文本流的实体关联关系抽取。例如在百度百科文本中会说明每个实体所属的类别。

纯自由文本。对于论坛上的各种自由文本数据，项目将基于文本处理技术的模式匹配方法，对实体之间的关联关系进行抽取。例如如果语句符合等顿模式（“苹果、香蕉、梨等水果”，则可以推导出苹果、香蕉、梨这三个实体是水果）。通过这种模式可以获得很多实体间的候选关系

(4) 语义消岐

该模块主要实现对文本文档抽取的实体和关联关系的语义消岐，在文本数据处理中将出现大量的同名异义和同义异名的现象。为了进行语义消岐，本项目将结合实体间的语义关联关系，利用实体网络上的聚类技术、基于 LSA 的语义距离计算等方法，分析两个实体对象是否指代同一个实体。

3.4 证券知识图谱构造

在证券知识图谱中将包含证券领域的主要概念、实体（包括金融产品、相关的企业、行业分类、主要人物、相关产品等）及其它们之间的关系，在知识图谱中还将包含节点类型等相关信息。

(1) 关联分析

该模块主要实现对来自多个数据源文本的综合分析，针对不同文本中抽取实体关系，构建实体关联网络，并基于实体关联网络分析实体之间的关联关系，借鉴异构信息网络分析技术，研究实体关联网络上实体间语义距离的度量方法，和面向大规模实体关联网络的距离计算方法。

(2) 信息融合

该模块主要实现对来自不同来源的实体和实体关联关系信息进行融合，本项目的数据主要来自于领域数据源、百科类知识文本和来自社交网络的自由文本的数据，这三类数据源形成的信息的可信度各不相同，本项目将结合实体和关联关系的语义关联，以及多数据源的不同的可信度，结合 D-S 证据理论，研究多源可信度评估模型。从而实现对多种数据上生成的实体和关系信息进行综合。

(3) 质量控制

该模块主要实现对知识图谱的质量评估和修复。由于知识图谱系统是自动构造的，所以其质量难以和有人工构造的知识库相比，为了提高证券知识图谱的应用效果，需要保证知识图谱始终处于一个比较高的质量。本项目将从如下两个方面展开研究：

研究基于众筹的质量控制体系。首先根据信息融合模块对实体及其之间关联关系可信度的评估，设计由工作人员回答的问题，在问题生成过程将偏重可信度比较低的知识片段。随后结合用户的回答情况和结果的可信度，结合回答问题的代价和总的成本，结合遗传算法等数据优化方法，以知识图谱系统的总体质量提升为目标，研究提问问题的选择算法。

基于规则数据补充方法，由于数据倾斜性的问题，我们生成的知识库往往缺失大量的信息。据我们评估，即使相 Probase 这种比较成熟的知识库，其中也包含了大量的错误信息。为此项目将对知识库中不同类型实体之间的关联关系模式进行挖掘，并将这些模式反过来应用于对现有知识的分析，以扩充知识库的容量。

(4) 动态更新

该模块主要实现对知识库的动态更新。由于数据抽取程序形成的事实容易收到错误文本数据的影响，所以需要筛选出那些永久成立的事实，抛弃在不可信数据源中出现的临时的事实。为此项目将结合文本处理模块生

成的候选实体和关联关系，研究迭代式的图谱更新方法。主要是研究增量式的知识图谱信息融合模型，该模型将建立基于时间属性的可信度评估方法，通过考虑时间衰减因子，分析该事实是永久事实还是临时事实。

3.5 社交网络分析系统

(1) 社团分析

该模块主要实现对微博等社交平台上社交网络上社团的分析，为了了解网络舆情，需要掌握网络上民众的情况，为此需要通过社团分析技术将网络上的用户或民众根据其标签属性（Profile）或用户之间的好友关系，进行社团分析。以将用户分成若干类，由于每一类中的用户在观念和行为上相对一致，所以可以通过对用户在网络上行为的分析，分析其观点的代表性。

基于社团分析的结果，系统还将进行用户画像，根据同一社团用户的相似性原则，对用户赋予标签，通过这一方式可以补充用户在定义自己的标签属性时所没有填写的内容，从而可以更准确地了解用户的性格和喜好。以便进行进一步准确地分析。

(2) 传播分析

该模块主要分析舆情在微博社交网络上的传播范围。该模块将根据社交网络的结构，以及历史上各种主题的信息在社交网络上的传播方式，生成各种主题消息在社交平台上的传播模式，并以此为基础，对各种消息在社交平台的传播范围进行预测。

(3) 内容分析

该模块主要实现对微博、论坛等社交平台上的短文本信息进行分析，该模块将对社交平台上的各种文本进行文档结构建模，然后利用生成模型等方法，结合文本的上下文，发帖人的历史行为、文本主题，文本的相关词集合（可以来自证券知识库系统）生成文本的主题模型。从而实现对短文本的分析

(4) 情感分析

该模块主要实现对文本的情感分析，项目将在金融知识图谱系统中建立情感词库，该情感词库将包含主要的情感词汇，以及这些词汇的情感倾向，并基于情感词库和不同社团的习惯，对文本的倾向进行分析。结合短文本的主题分析技术，获取面向特定主题的情感。

3.6 情绪分析

(1) 模型定义

该模块主要实现对情感指数模型的设计，该模型将主要考虑近期市场交易走势（包括单个金融产品的价格、成交量、资金流向等以及金融产品群体的共同变化趋势）、社交网络和论坛上舆论的情感倾向、近期的主要公告内容倾向等众多因素，并结合人工标注信息利用统计模型设计情绪指数模型。

(2) 行情分类

该模块主要实现对证券行情的分类，在分类中将结合价格和成交量因素，利用划线方法对价格和成交量波动情况进行分割，使得在一定时间区间内，保持同样的价格和成交量变化趋势。然后，结合序列相似性度量方法通过聚类方法将相似的序列片段聚在一起，通过对聚类的语义描述实现对行情的分类说明。行情的分类将作为情绪指数的重要因素。

(3) 参数训练

该模块主要实现对指数模型中参数的训练，在指数模型中包含众多参数，本模块将结合线性回归等方法，利用大量历史数据对模型的参数进行训练。以获取具有较高区分度的参数配置。

(4) 情感计算

该模块主要实现对情绪的实时计算，项目将搭建基于云计算平台和流计算模型的情绪计算系统。每天通过网络爬虫系统从网络上采集相关的数据，并通过文本处理模块和社交网络分析系统实时计算舆情状态，并通过对价格和成交量序列的计算，获取行情分类情况，最终实现对情绪的计算。

4 关键技术

4.1 图谱构建知识采集

4.1.1 大规模证券领域知识资源获取

为了创建证券知识图谱，我们将对网络空间中的证券领域知识资源进行采集获取，需要涵盖在线百科、领域网站、微博等多种知识发布与交换通道，需要能够建模网页主题，评价网页主题相关度并能对网页主题进行预测，需要能够提供实时、增量、鲁棒的采集系统架构，因此拟从如下几个方面研究数据的特色采集技术

(1) 基于页面内容和 URL 分析的主题相关度评价和预测算法: 研究基于网页内容（文本、数据等资源）特征的网页主题相关度计算方法和基于 URL 自身携带的信息、锚文本等信息对 URL 所指的页面进行主题相关度预测算法

(2) 大规模、分布式高效并行采集与数据存储框架：针对海量数据的采集与数据存储，设计实现分布式的任务执行框架。支持超过 10 万个采集探针和数据存储节点同时工作，能够保证覆盖各个采集通道上国际、国内主流信息发布服务。

(3) 增量采集技术保证数据实时性：互联网络中信息量的快速增长使得增量采集技术成为网上信息获取的一种不可或缺的手段，可以避免重复采集未增变化的网页带来的时间和资源上的浪费。我们利用 J. Cho 等定义的网页的时新程度：估算单个网页在一段时间内的平均时新性及平均年龄并以此估计网页的变化，实现网页的增量采集。经实际系统验证，该方案实际可操作，降低了信息的冗余度，提高了采集效率。

4.1.2 面向证券领域的深网资源采集方案

为了获取证券论坛中的数据，需研究面向证券领域的深网数据采集。区别于互联网中搜索引擎可见的、易获取的浅网(Surface Web)信息而出现的。私有/受限网页（需登录才可访问的网页）、表单结果和脚本形式呈现的内容，等都属于深网的范畴。深网采集技术是采集领域的传统难题。

拟研究一种面向证券领域（应用）的基于抽样的深网采集方法，该方法采用人工辅助的方式，利用人工定义的资源采集的需求描述自动实现领域深网资源采集。针对深网资源采集，拟从多方面展开科研攻关。例如，对表单结果，将根据领域或应用需求描述，实现领域知识抽样（采集任务抽样），然后根据领域知识样本实现指定站点的深网资源采集；我们对微博客的 API 受限的情况，将使用多代理、多渠道分散采集的方法，打散采集请求，提高单位时间的采集能力。又如，对评论、视频源地址等脚本信息的采集，我们拟通过全自动模拟浏览器用户上的行为，能够稳定地对各种脚本信息实施异步、分布式的并行高速采集。

4.2 知识抽取

4.2.1 直接抽取知识

通过上一步知识采集方法能获得海量的互联网页面。这些页面中往往存在着许多结构化的知识，这些知识以非常规律的方式展现在用户面前。如图 4 所示，图为百度百科对于“货币”这一证券术语的解释。黑色边框部分为结构化数据。用于描述“货币”这一术语的主要特征。

这类结构化数据的存在，使得用户能在第一时间获得关于该术语的相关知识介绍。通过分析页面结构特性，可通过基于模板的匹配方法进行知识抽取。这类知识由于是由人工精心编辑而成，故质量有所保证。

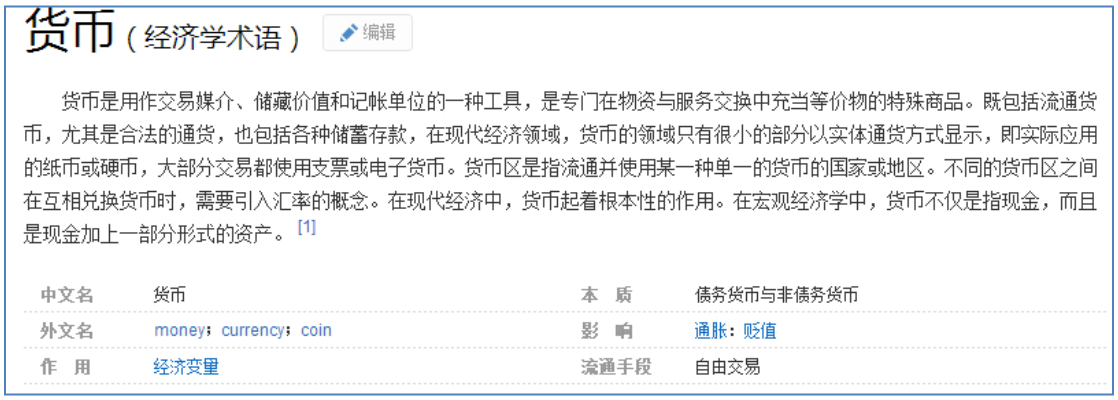

图 4 网页中的结构化数据示例

4.2.2 从文本中抽取知识

互联网页面中更广泛存在的结构为非结构化文本。要从这类数据中抽取知识，基本思路为将从文本中抽取新知识建模为数据挖掘的分类问题。

分类器用来判断句子中的哪个位置上的单词为关系中的实体。因此，在构建训练集过程中，需要构建关系与句子的对应关系以及关系实体与句子位置的对应关系。特征通过对句子进行语法解析以及词法解析获得，通过最大熵模型来判断关系的分类，同时利用条件随机场模型来判断实体所在句子的位置。如句子“国债种类有凭证式国债、实物式国债、记账式国债三种”。首先通过第一类分类器可以得知，该句子是在描述实体“国债”的“种类”属性，然后通过第二类分类器，获知“凭证式国债”、“实物式国债”以及“记账式国债”均为“种类”属性的值。

4.3 观点挖掘

针对证券类评论信息，研究评价对象抽取与情感倾向挖掘。

4.3.1 评价对象抽取

使用基于句法分析的评价对象抽取技术。对于给定语料，首先对其进行分词、词性标注以及句法分析等处

理，然后提取其中的名词(NN)和名词短语(NP)得到候选评价对象；继而对候选评价对象使用频率过滤、PMI 算法和名词剪枝等算法进行筛选得到最终的评价对象表。

候选评价对象抽取。评价对象可以是名词或名词短语。关联规则挖掘方法提供了一种抽取语料中频繁项的方法，可以有效的抽取语料中的评价对象。但其存在一定问题，就是在判定两个词能否形成短语时仅考虑了两个单词的共现次数，并没有考虑到短语的句法结构，为此，我们引入了句法分析技术来抽取评价对象。考察情感句“这个证券很有前途，市值很高。”使用关联规则挖掘方法，若“证券”和“市值”的共现次数足够多，则“证券市值”会被识别成评价对象。按照我们的句法限制，此句的评价对象为“证券”、“市值”，可以看出，通过使用句法信息，我们能够避免抽取出“证券市值”这种错误评价对象，但同时也会产生一定的噪声(续航、能力)，为此，我们引入了三种过滤技术，下面对其进行详细介绍。

4.3.2 评价对象筛选

由句法分析得到的候选评价对象集存在一定的噪声，为此，我们加入了相应的过滤技术。评价对象筛选流程如下：首先，使用词频信息进行过滤；其次，使用基于网络挖掘的 PMI 算法进行过滤；最后，使用名词剪枝技术解决单个词的冗余现象。下面分别对这三项技术进行介绍：

(1)词频过滤：

词频过滤即将语料中出现次数比较少的 NN 或 NP 过滤掉。词频信息过滤的加人主要基于两点考虑：1. 评价对象更倾向于在评论中多次出现，一些不相关的 NN 或 NP 应该在商品中很少出现，如“有限公司”、“多图”等。2. 词频信息过滤可能会过滤掉

一些评价对象，但这对系统的结果影响不会很大，因为那些出现次数较少的评价对象并不被大多数人所关心，属于次要属性。

(2)PMI 过滤(Point wise Mutual Information)：

PMI 值能够量化词与词之间的关系，在一定的文本集合中，词 a 和 6 的 PMI 值定义如下：PM 一一 N / (N。×N)其中，N 曲表示既包含 a 又包含 b 的文本数，N。表示含有 a 的文本数，N 表示含有 b 的文本数。从公式中可以看出，PMI 值的计算使用了统计的思想，同时基于这样一个假设：两个单词共现的次数越多，则它们之间的联系也就越大。PMI 值计算的难点在于大规模文本集合的获取，理论上讲，文本数越多，则统计效果越明显，PMI 值的计算也应该越准确。使用 PMI 值来进一步挖掘评价对象的领域相关性。为了得到足够大的语料，选取百度的搜索结果作为语料库。方法如下：针对每一领域，选取最具代表性的词语 Y,U，计算候选评价对象侧与相关领域 W 的 PMI 值，值越大，则说明的相关性越强，更可能成为一个评价对象。最后通过设定阈值的方法进行过滤，实验表明这种方法取得了比较好的效果。

(3)名词剪枝：

此技术主要应用于冗余名词的过滤。为了说明什么是冗余，首先定义 s-support：对于名词 t，设包含 t 的句子数为 S，在 S 个句子中，f 单独作为评价对象出现的句子数为忌(这是个句子中不含有包含 t 的短语)，则 s-support—k / s。对于 s-support 值小于 0.5 的名词评价对象，认为它是冗余的，作过滤处理。

4.3.3 情感分析

针对该任务，首先分析情感句的结构，将其分为四类；继而针对各类制定相应的倾向性判断规则，最终基于无指导的方法完成评价对象的倾向性判断。下面对其进行详细介绍。

(1)情感句型分类

通过观察语料，将句子分为四类，具体定义如下：

类别一：句子带有明显的倾向性，即句子中带有一种倾向性(褒义或贬义)的上下文无关情感词明显多于另一种。例如：“这个证券公司很有潜力”，这句话的情感词为“有潜力”，为褒义，则此句明显含有褒义的倾向性。

类别二：句子不带有明显的倾向性，但句子中含有情感词，且褒义和贬义情感词的个数相同。例如：“这个证券公司市值很低但有潜力”，这句话中含有“很低”和“有潜力”两个情感词，极性分别为贬义和褒义，但不能说此句的倾向性是褒义还是贬义。

类别三：句子不带有明显的倾向性，且句子中没有情感词，但其上下文的句子带有明显倾向性。例如：“这个证券公司市值很有潜力，公司正在发展”，“公司正在发展”这句话本身不含有情感词，但此句的前一个句子“这个证券公司市值很有潜力”带有明显的倾向性。

类别四：句子不带有明显的倾向性，句子中没有情感词，且其上下文的句子也不带有明显倾向性。例如：“投资证券要关注证券公司。”。

(2)倾向性判定规则

针对前三类句子，分别定义规则对其进行处理，这三个规则也是倾向性判断的基础；对于第四类句子来说，无法找到一个通用的规则，为此引入上下文相关极性词的概念，下面对具体规则进行详细介绍：

规则一：对于第一类句子，句子中评价对象的极性与句子的极性相同，如“这个证券公司很有潜力”，这个句子的倾向性为褒义，则其中的评价对象“证券公司”为褒义。

规则二：对于第二类句子，找出句子中的评价对象，针对每个评价对象，选取评价对象 8 个字(实验表明这个窗口大小比较合适)的范围内与其最近的情感词作为直接修饰它的情感词，评价对象的极性即与修饰它的

情感词的极性相同。

规则三：对于第三类句子，使用上下文信息进行判断，当前句子优先与前一个句子的极性相同，如其前一个句子也不存在明显的倾向性，则认为当前句子与其后一个句子的倾向性相同。句子的倾向性判定之后，则句子中所有评价对象的极性与此句子的极性相同。

规则四：最后一类句子中包含了剩余的所有句子，为了进一步挖掘此类句子中的情感句，引入了上下文相关情感词的概念。考虑“这个证券公司市值高。”这个句子，“高”这个词本身并不含有极性，但在修饰“市值”的时候，就含有了一定的极性，我们将这种词称为上下文相关情感词。针对这种情况，相应的评价对象倾向性判定方法如下：首先找出句子中的评价对象，然后查找距离其最近的上下文相关情感词，如果二元对<情感词，评价对象>含有极性，则抽取结果，否则过滤掉。上下文相关情感词表较小，并搭配评价对象而产生倾向性，该词表由人工构建。这四类句子的处理优先级与上面介绍的顺序相同，即对一个句子，依次判断它是否属于某一类，属于则按照相关规则处理，给出倾向性结果，若无倾向性，则直接过滤掉。

5 小结

本文介绍了知识图谱概念，以及证券行业知识图谱的特点及应用前景，并对建设证券行业知识图谱展开了若干相关关键技术研究。具体来说，主要研究了系统架构与主要模块，包括数据源、网络爬虫系统、文本分析系统、证券知识图谱构造、社交网络分析系统、情绪指数分析等技术方案，并就图谱构建知识采集、知识抽取和观点挖掘等关键技术展开了深入讨论和研究。

知识图谱的构建是跨学科的结合，需要行业知识库、自然语言处理，机器学习和数据挖掘等多方面知识的融合。作为证券行业文本挖掘的重要基础工作，建立起证券领域全行业的知识图谱，将有助于实现行业中自动化低延迟事件关联、作用链关联、产业链关联，从而为证券市场的行为提供关键讯息支持和交互印证，这将大大提高互联网文本数据对于证券行业的价值和判断的准确性。作为证券文本语义理解和语义搜索的关键基础技术，构建证券领域知识图谱为未来证券领域文本分析、舆情监控、知识发现、模式挖掘等提供了坚实支撑。

参考文献：

[1] http://www.keenage.com/html/c_index.html

[2]梅家驹等编. 同义词词林，上海辞书出版社，1996.

[3] 黄曾阳. HNC 理论概要. 中文信息学报，1997.

[4] http://36kr.com/p/173903.html