基于大数据分析和复杂事件处理的金融信息服务平台①

刘 斌

(兴业证券股份有限公司, 福州 350001)

摘 要: 针对大数据时代下金融信息服务滞后性、片面性、同质化的问题, 设计并实现了基于大数据分析和复杂事件处理的金融信息服务平台. 该平台采用多通道采集、浅层语义处理等技术实现多源数据的采集、抽取与清洗, 采用面向金融领域特征的网络观点分析等技术进行数据分析, 为证券投资者、投资顾问及机构等三类用户提供及时、精准、个性化的金融信息服务, 取得了良好的应用成效.

关键词: 大数据分析; 复杂事件处理; 情绪分析; 数据采集; 金融信息服务平台

Financial Information Service Platform Based on Big Data Analysis and Complex Event Processing

LIU Bin

(Industrial Securities Co. Ltd., Fuzhou 350001, China)

Abstract: In view of the problems of lag, one-sided and homogeneous financial information service in big data era, this paper designs and implements a financial information service platform based on big data analysis and complex event processing. The multi-source data acquisition, extraction and cleaning are implemented by multi-channel data acquisition and shallow semantic processing technology on the platform. And the data is analyzed by the financial network sentiment analysis technology. The platform provides timely, accurate, personalized financial information services for securities investors, investment adviser and institutions and achieves good application results.

Key words: big data analysis; complex event processing; sentiment analysis; data acquisition; financial information service platform

1 前言

大数据时代的金融信息呈现海量、异构等特点, 广大投资者或金融信息的使用人员无所适从. 因此, 对金融信息服务在及时性、精准性、差异性等方面提出了更高的要求. 如:

① 金融信息的发布具有时间不确定、渠道多样化、数量巨大等特点, 投资研究人员花费大量的时间在海量数据整理中, 无法专注于核心工作, 效率低下, 希望有更高效的工具可以及时、系统地为其提供所关注方面的信息;

② 互联网已经成为机构、上市公司以及投资者信息发布与获取的重要渠道, 政策法规、公司公告、热门事件、投资心得、自媒体等海量异构数据以及各种噪声信息使得传统的人工模式已很难从中精确地获取最有价值的信息; 市场的一体化使得信息与事件不再孤立, 而传统金融信息服务只是向客户提供客观资讯或事件, 缺乏关联性的分析, 造成价值丢失. 对于广大的投资者来说, 数据的分析与处理专业性强、条件复杂、门槛高、成本大;

③ 互联网的开放性与随意性使得金融信息更趋向同质化. 而投资者更关心的是与自身投资相关的金融信息、账户信息、资产变动等相对个性化的信息服务; 证券公司等机构为了防止恶意及虚假的信息影响公司品牌形象, 防范舆论风险、市场风险, 需要更实时更全面发现对机构本身会产生影响的互联网信息风险点.

本文基于上述金融信息服务的新要求, 设计并实现了基于大数据分析和复杂事件处理的金融信息服务平台, 同时还展开介绍了数据采集、抽取和分析等关

① 收稿时间:2016-07-15;收到修改稿时间:2016-08-29 [doi:10.15888/j.cnki.csa.005706]

键技术, 并描述了系统的功能. 为三类用户能提供精准、及时、个性化的金融信息服务, 取得了良好的应用效果.

2　设计与实现

2.1 系统架构设计

针对数据源多源异构、信息服务实时个性化的需求, 新一代金融信息服务平台架构应达到如下要求: 数据源方面, 能准确地获取多源数据, 并能对结构化数据以及非结构化数据进行清洗并统一存储; 在分析方面, 既能对海量历史数据进行批处理, 也要能实时处理流数据; 在信息发布方面, 能针对不同的用户实现统一、标准、及时的个性化信息发布. 系统在架构上主要从以下几个方面进行分析与设计.

2.1.1 数据源方面

本平台需要获取的数据来源包含: 互联网及社交媒体、金融资讯中心、客户数据中心等. 互联网数据主要涉及交易所、央行、证监会、主流财经网站等公告及新闻; 社交媒体包括微博、微信、股吧等; 金融资讯中心主要包括如个股行情、大盘指数、行业新闻、研究报告、公司信息、市场数据等; 客户数据中心主要包含如客户资料、持仓信息、交易流水、行为日志等.

2.1.2 信息处理方面

要实现对多数据源的采集、抽取和标准化, 并具备对多数据源协同分析的能力, 能从多个维度对海量异构数据进行实时处理分析. 要求本平台具备有一个能实时处理信息的引擎, 实现对实时事件的处理, 提供有效的金融信息, 提升投资顾问和投资者及时准确掌握和利用市场信息的能力;

为满足投资者对金融信息差异性的需求, 要求平台能对投资者进行分类分析, 得到投资者的个性化需求, 才能针对性地投资者提供差异化的金融信息;

为提升数据价值, 满足金融信息精准性的需求, 要求平台能对多源数据进行有效挖掘, 构建数据的分析模型, 如热点主题、投资者情绪指数、股市预测等投资者感兴趣的模型, 获取数据的有效价值, 提升金融信息质量, 提升平台服务水平等.

2.1.3 信息发布方面

为有效地将金融信息实时差异化地推送到投资者, 需要本平台能整合各信息发布终端, 打通各个渠道, 实现金融信息的统一发布平台, 为证券投资者及投资顾问提供全面及时的信息服务.

通过以上对系统架构的分析, 本文设计了基于大数据分析的金融信息服务平台系统架构, 如图 1 所示, 主要包含数据获取层、数据分析层和数据应用层.

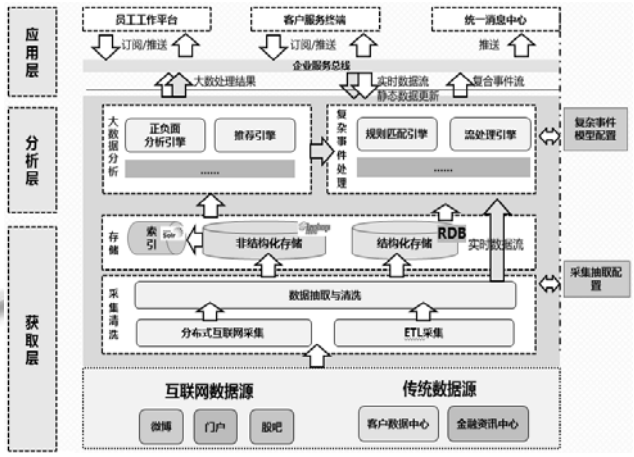

图 1　平台系统架构

2.2 系统实现

基于对系统架构的设计要求, 分别对系统的数据获取层、数据分析层以及数据应用层进行实现.

2.2.1 数据获取层

数据获取层负责多源异构数据的快速获取、清洗、存储, 如图 2 所示. 通过“基于多通道技术”实现互联网数据的分布式统一采集; 通过“基于浅层语义的网页抽取技术”实现海量混杂数据的统一抽取与清洗; 利用 MySQL 数据库及 HDFS 分布式文件系统实现结构化、非结构化海量数据的存储; 利用 SolrCloud 实现高效全文索引.

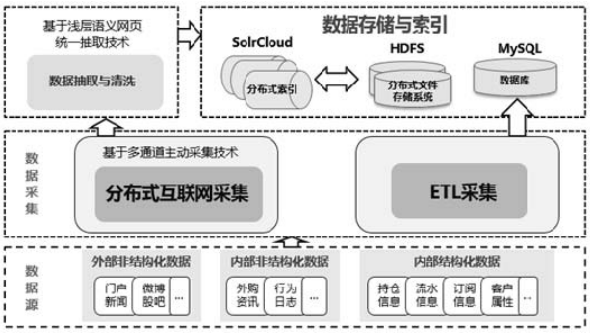

图 2　数据获取

数据获取包含传统数据源和互联网数据源的获取. 数据获取是否正确直接影响后续的数据分析及应用.

因而, 针对传统数据源, 主要为内部数据, 在采集过程中通过内部数据校验机制对采集的数据结果进行验证审核; 针对互联网数据源, 主要为外部数据, 通过定时监控结合人工审核的机制对获取到的数据进行验证, 从而保证了内外部数据获取的准确性.

2.2.1.1 传统数据获取

针对传统数据源, 主要是金融资讯中心和客户数据中心的内部结构化数据, 采用传统 ETL 工具, 从源端将数据采集到目标数据库中.

2.2.1.2 互联网数据获取

针对互联网数据源, 采用如下三个模块实现数据获取:

1) 互联网采集模块

利用网页采集工具, 针对指定的页面和主题进行采集, 并以网页的形式下载到本地. 信息采集模块是系统的基础模块, 所采集的网页是后续模块的输入数据. 采集模块应包括微博采集、门户采集、公告采集, 实现对微博数据、重要门户网站数据以及公告新闻数据的采集.

2) 数据抽取模块

对采集模块得到的网页进行清洗预处理, 去除页面结构错误. 通过算法定位到抽取内容的标签节点, 抽取出标题、正文、时间等所需的信息, 将其存入数据库并生成 XML 文件.

3) 索引模块

对抽取得到的 XML 文件, 根据自定义的索引规则, 将 XML 文件信息进行关键字提取, 设置标签, 然后将文件加入到索引库中, 以供后续检索与分析功能使用. 索引建立的过程, 类似于将数据进行关键字提取, 设置标签, 在后续工作中, 可以通过这个标签进行内容过滤获取期望数据的操作. 归结起来大致的过程为: 获取数据→设置建立索引规则→建立索引→写入磁盘/内存.

在互联网数据获取方面, 多源、异构数据的统一采集、抽取与清洗是该环节的关键点和难点, 本文采用“基于多通道主动采集技术”实现互联网数据的分布式统一采集, 研发“基于浅层语义的网页抽取技术”实现海量混杂数据的统一抽取与清洗.

2.2.1.3 基于多通道的主动采集技术

该技术分为非常规采集和常规采集:

1) 非常规采集

非常规采集共分为四个部分: 任务分发器、Cookie 生成器、主题与种子 URL 定制、非常规采集器, 如图 3.

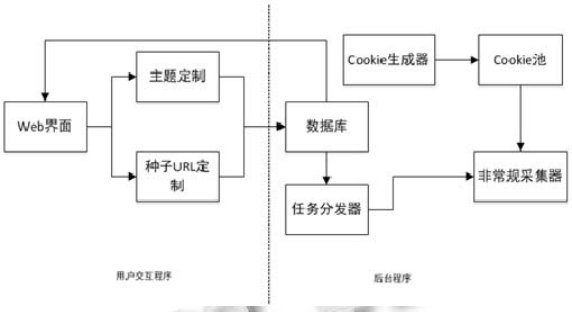

图 3 非常规采集结构图

任务分发器负责将需要采集的页面及其相关信息整合, 并根据优先策略分发给微博采集器. Cookie 生成器为后续的页面下载模块提供登录 Cookie, 是页面下载模块的基础. 为了简化使用的复杂性, 该模块主要应用在系统部署时. 主题与种子 URL 定制模块提供用户设置采集目标的界面, 采集目标包括关键词和用户主页的 URL. 用户可以通过 Web 用户界面, 设置关键词, 从而采集相关的内容, 也可以设置 URL 采集目标用户发布的内容. 关键词与 URL 均存放于后台数据库中. 非常规采集器提供页面下载功能. 可供设置的内容包括采集间隔时间与每次采集并发线程数, 根据不同媒介、以及贷款设置相对合适的采集间隔时间与并发线程数. 主要面向系统部署人员, 以参数的形式输入给程序. 程序将以在系统后台运行. 在终端运行程序时输入参数, 包括: 输出目录、采集间隔时间、并发线程数. 输出目录产生与关键词和目标用户主页的页面文件.

2) 常规采集

常规采集主要是指静态网页数据采集, 主要由网页采集模块、链接抽取模块和链接判重模块三部分组成. 静态网页数据采集是从一个初始链接对应的网页开始采集该网页的源代码, 并且在保存网页源代码的同时, 不断地从中抽取出新的链接. 程序重复上述过程, 直到满足采集深度达到事先设定的值或者链接集合为空. 其系统流程图如图 4 所示.

网页采集模块实现获取链接对应网页源代码, 并将网页源代码保存到文件中. 链接抽取模块抽取网页源代码中的链接和对应的锚文本, 并保存链接和对应的锚文本信息在指定文件中. 链接判重模块可以初始

化一个集合，可以往集合中添加链接元素，并判断某个链接是否在集合中.

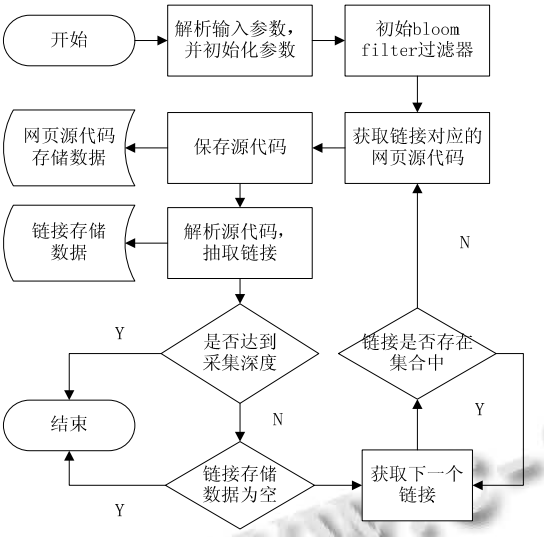

图 4　静态网页数据采集系统流程图

2.2.1.4 基于浅层语义的网页抽取技术

基于浅层语义的网页抽取技术主要分为长文本网页抽取和短文本抽取:

1) 长文本抽取

长文本网页抽取主要由四个模块组成，如图 5 所示.

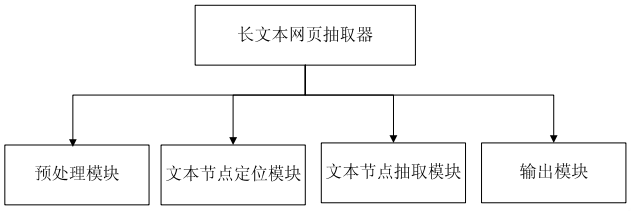

图 5　长文本网页抽取模块图

预处理模块对网页源码预处理，过滤掉噪声标签节点并对网页源码中错误地方进行修正，如标签匹配错误等. 文本节点定位模块根据预处理模块提供的 DOM 树结构，通过计算节点文本密度，在所述 DOM 树中来定位正文区域. 文本节点抽取模块根据文本节点定位模块提供的正文标签节点按照先序遍历 DOM 树结构中正文节点子树，抽取出遍历过程中各节点的文本内容. 输出模块用于检查抽取的文本是否符合条件，将抽取好的正文和属性作为输出项存储到数据库和文件中.

2) 短文本抽取

短文本抽取即为多记录网页抽取，主要由四个部分组成，如图 6 所示.

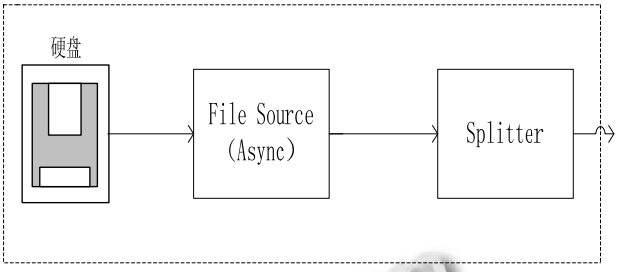

图 6　多记录网页抽取模块图

预处理模块提供抽取过程所需的 DOM 树结构. 记录区域定位模块根据预处理模块提供的 DOM 树结构利用横向层次分析法在 DOM 树中来定位记录区域. 记录分隔符识别模块根据记录区域定位模块提供的记录子树利用双向搜索方法从记录区域块中找到记录之间的分隔符并进行存储. 输出模块根据记录区域定位模块提供的记录子树和记录分隔符识别模块提供的分隔符先序遍历记录子树并输出到文件.

2.2.2 数据分析

数据获取完毕后进行数据分析工作，数据分析层包含大数据分析部分和复杂事件处理部分，主要负责海量数据批处理及实时流数据分析.

2.2.2.1 大数据分析

在大数据分析方面，采用“融合用户观点和用户行为的证券应用技术”、“面向证券领域特征的网络观点分析技术”、“分/聚类技术”等主题分析、情绪分析以及投资者分析工作，如图 7 所示.

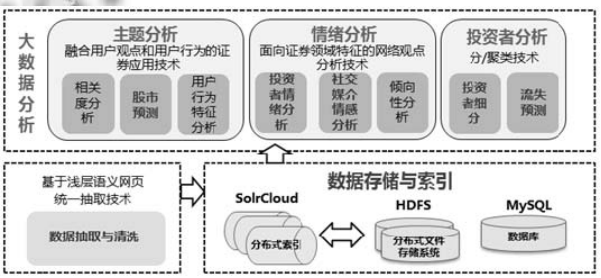

图 7　大数据分析

1) 主题分析

① 相关度分析: 根据用户自定义的主题及主题关键字，计算新闻与主题的相关度值，并将相关度值高于阈值的新闻展示出来，提供给用户浏览.

② 股市预测: 根据社交媒介情感分析量化结果，

感知市场情绪，并构建股市预测模型，预测股指判断市场走势.

③ 用户行为特征分析：在用户登陆系统后，通过对用户显式或隐式采集到的行为，分析出用户的行为特征，并以此为依据，将用户可能感兴趣的证券信息推荐给用户.

2) 情绪分析

① 投资者情绪分析：根据互联网以及行业数据构建投资者情绪指数模型，感知投资者情绪，辅助投资决策.

② 社交媒介情感分析：根据社交媒介用户发表的内容以及社交关系，对用户发表的内容进行情感分析，得到社交媒介对某一类事物的观点倾向.

③ 倾向性分析：根据新闻与主题关键字，判断该主题下新闻的倾向性，并分别置为-1(负面)，0(客观)，1(正面).

3) 投资者分析

① 投资者细分：获取投资者的行为数据进行分析，建立投资者细分模型，根据投资者的行为特征将投资者进行划分.

② 投资者流失预测：监测投资者的行为数据，建立投资者流失预测模型，识别投资者流失倾向.

通过上述的主题分析、情绪分析、投资者分析，整合互联网数据以及内部数据，挖掘数据的有用信息，从而将其推送给用户或投资顾问.

在数据分析中，证券领域的情感量化是情绪分析的关键点. 本文采用基于异构图模型的证券情感量化技术用以解决情感量化问题.

首先对要进行情感量化的文档进行预处理，然后基于证券倾向性文档以及证券情感词构建二分连接图，计算证券情感词的倾向性权重，最后利用文档的相关性得分以及倾向性得分根据概率模型计算文档的情感得分. 该方法的核心关键是计算证券情感词的权重，为了获取证券领域的情感倾向，在利用该方法进行帖子的情感量化时，所基于的倾向性文档集采用了证券领域带有倾向性的文档. 证券情感量化具体过程按如下步骤进行:

1) 预处理

对证券倾向性文档进行预处理，包括去除标点符号、网页链接、表情符号、特殊符号等噪声，分词，去除停用词.

2) 基于异构图的证券情感词赋权

为了获取证券领域特定情感词的权重，在如下构建的二分连接图中，其倾向性文档均采用带有证券倾向性情感的文档集合，使用如下方法计算证券情感词权重.

在由证券领域的倾向性文档集和证券情感词组成的二分连接图，如图 8 所示.

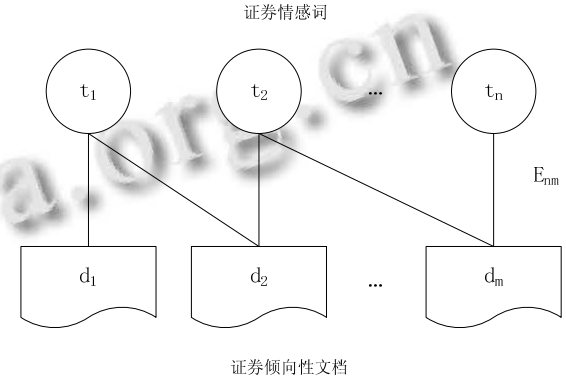

图 8 倾向性文档-倾向词二分连接图

根据式(1)和式(2)计算倾向性文档和证券情感词每步迭代的得分.

$$AuthScore^{(T+1)}(d_j)=\sum_{i=1}^{n} w_{ij}\times HubScore^{(T)}(t_i) \tag{1}$$

$$HubScore^{(T+1)}(t_i)=\sum_{j=1}^{m} w_{ij}\times AuthScore^{(T)}(d_j) \tag{2}$$

当连续两步迭代的情感词得分和倾向性文档得分误差小于某一阀值，迭代计算结束，得到最终的每个情感词得分即为每个证券情感词的倾向性权重.

3) 情感量化

根据式(3)计算倾向性得分，结合倾向性得分并根据式(4)计算最终的情感得分.

$$Score_{opn}(d)=(1-\lambda)\sum_{t_i\in d} a_i\frac{co(t_i,q)}{c(p,d)\cdot len(d)}+\lambda\sum_i a_i\frac{\sum_d co(t_i,q)}{\sum_i\sum_d co(t_i,q)} \tag{3}$$

$$RankScore(d,q)=Score_{rel}(d,q)*Score_{opn}(d,q) \tag{4}$$

2.2.2.2 复杂事件处理

复杂事件处理主要负责实时事件流的处理及不同事件实时关联分析. 主要包括：事件输入、事件处理及事件响应三部分，如图 9 所示.

事件输入来源包含大数据分析结果及实时数据流，数据覆盖全面实时；在事件处理部分，研发可视化规则定义和基于 EPL 的事件模型定义,可以方便用户自

主定制事件模型及业务规则，并利用热切换技术实现模型在线发布；在此基础上，根据业务需求研发针对客户应用及员工应用的复杂事件处理模型集；事件响应部分负责将事件处理的结果通过输出适配器应用于客户及员工系统.

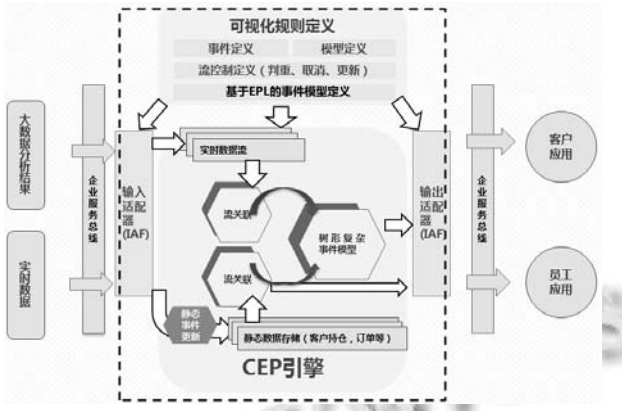

图 9　复杂事件处理

复杂事件处理的基础就是事件间的关系. 其中，事件之间的偏序由因果关系、时间关系决定，而一个事件对一个事件集合的总结、代表或指示关系则是组合关系. 本平台通过实时复杂事件引擎的搭建和多输入多输出个性化服务模型的构建实现复杂事件处理:

1) 实时复杂事件引擎的搭建

① 实现以事件驱动为核心的服务模式，采用内存处理技术，并通过索引化流事件查询规则，实现对实时流事件的高效分析;

② 结合企业服务总线实时事件发布(ADB)与复杂事件处理实现基于可变滑动窗口的静态数据与实时事件关联技术构建海量静态数据的实时维护解决方案，提升了海量静态数据处理的实时性;

③ 通过自定义标准化底层事件处理协定，实现了事件的过滤、判重与取消，降低了事件流的复杂度，提升了核心处理模块的处理效率.

2) 多输入多输出个性化服务模型的构建

① 多源异构事件构成的复杂事件处理引擎的“多输入”.

平台通过事件适配层(IAF)对接 EMS 消息队列接口，订阅不同来源实时发布的事件，包含公司数据中心，资讯中心，以及大数据分析系统等. 目前系统定义的“多输入”事件包含: 客户特征数据(客户关键时点事件、客户风险偏好、客户满意度、客户贡献度等); 客户行为数据(买卖流水、银证转账流水、终端访问日志); 市场数据(实时行情特征数据、资讯数据等); 大数据分析结果(个股特征数据、市场热点，市场情绪分析等等数据). 这些“多输入”事件在复杂事件引擎中被定义为一个元事件及其流监听.

② 基于事件流关联的复杂事件处理模型构建.

在上述“多输入”的元事件基础上，可根据客户订单，业务分析等方式，定义出有特定意义的复合事件监听模型.

2.2.3 数据应用层

在数据应用层，通过企业服务总线(ESB)集成客户信息、员工信息及统一消息服务，为客户及员工提供全面及时的信息服务，如图 10 所示.

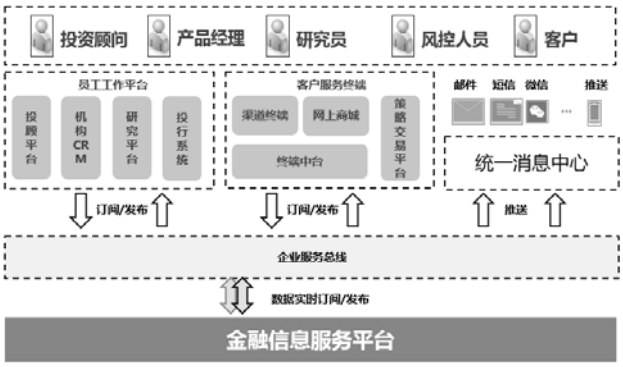

图 10　数据应用

秉承 SOA 理念与企业的整体 IT 规划，遵循统一标准，通过企业服务总线与各信息系统进行松耦合整合. 各系统包括大数据挖掘分析产生的事件通过企业服务总线进入复杂事件引擎，经事件处理模型产生的结果事件也是通过 ESB 提供给各应用终端送达用户.

平台提供的数据应用按照用户角度分为两类:

1) 客户类: 包括 PC 终端、移动终端、中台、网上商城、短信、邮件平台等与客户服务相关的终端系统，直接为终端客户提供个性化的实时证券信息服务;

2) 员工类: 包括投资顾问平台、机构 CRM 平台、研究平台等与员工工作平台相关的终端系统，为员工进行产品研究、市场分析、客户服务等提供全面、及时、便捷的证券信息服务.

3　功能、特色及成效

3.1 系统主要功能

本平台主要为三类客户提供金融信息，如表 1 所

示, 主要包含市场研判、即时资讯、风险监测、专题资讯、账户提醒、行情预警等六个方面的服务内容.

表 1 资讯服务内容

市场研判	融合各类信息进行分析作出市场判断	投资者情绪指数
		热点话题聚焦
		热股聚焦
		优理宝策略
		今日风向标
		盘前赢家
即时资讯	实时资讯、公告提示	业绩预告
		业绩快报
		分红提示
		退市提示
		配股通知
		公告提醒
		增发提示
风险监测	获取反馈信息及时提示风险	融资融券担保比例提醒
		约定式购回履约保障比例提醒
		重仓股重要新闻提醒
		资产异动提醒
专题资讯	跟踪企业机构重要信息	宏观时政
		行业要闻
		同业新闻
		企业重大事件跟踪
账户提醒	账户重要变动信息提醒	新股中签提醒
		客户开户首月无资金转入提醒
		客户账户大额亏损提醒
		客户大额赎回提醒
		持仓/自选股重要新闻提醒
行情预警	行情变化通知	个股到价
		指数到价
		大盘异动

为了保证金融信息服务的及时、精准、差异化的要求, 在信息获取方面, 在数据源分类分级的基础上, 通过系统自动处理和人工审核相结合的运营机制保证信息的及时性和准确性; 在分析方面, 基于大量的历史数据采用有效的挖掘分析方法进行训练, 在投资者分析模型、投资者情绪指数、股市预测模型等构建方面在稳定性及预测精度方面均有较大提升; 在信息推送方面, 基于复杂事件处理技术, 结合投资者分析模型, 能够实现金融信息流的即时、个性化的推送.

本平台支持多种渠道为投资者、投资顾问及机构提供及时、个性化、精准的金融信息服务. 通过手机终端、微信公众号等渠道为客户提供个性化资讯服务; 通过投顾工作平台、机构 CRM 等员工工作平台服务于公司投资顾问、资管产品经理、机构客户经理、行业研究员等, 有效提升员工的金融信息服务能力; 通过专题资讯分析, 服务公司董秘处、中高层领导, 辅助公司管理决策, 维护品牌形象.

3.2 平台特色及成效

金融信息服务平台为解决互联网数据因泛在分布、动态化、多样化等特点而难以采集的问题, 研发了一套针对互联网门户、股吧、微博、微信等自媒体资讯的分布式统一采集平台, 提高了采集精度. 为了实现在海量异构数据中分析和提炼准确有效的量化信息, 利用金融领域特征的网络观点分析、大数据文本挖掘等关键技术, 提升挖掘精度及数据价值. 同时构建实时个性化服务引擎提供实时差异化服务, 开创金融信息服务的实时处理新模式.

本平台在上述特点的支撑下提供及时、精准、差异化的金融信息服务, 从而有效提升投资者服务体验、提高员工专业服务能力、提升机构用户的公司信息管理水平, 应用成效归纳如下:

① 差异化、高质量的金融信息服务有效提升了投资者服务体验.

本平台利用个性化推荐技术将用户感兴趣的金融信息精准地推送给用户, 实现降低资讯服务通道成本的同时提升用户体验. 同时, 平台通过综合互联网数据、市场交易数据、客户数据等, 并基于大数据文本挖掘技术融合各类数据进行分析, 提供诸如投资者情绪指数、股市趋势预判等大数据投资者服务信息, 不仅丰富了金融信息服务的内容, 同时提升了金融信息的质量.

② 专业化的金融信息服务平台有效提高了员工专业服务能力.

本平台通过市场研判(热点话题、热点个股、股市趋势预判等)、即时资讯、风险监测等服务, 将重要资讯信息以及相关提醒信息推送给服务人员, 辅助服务人员将及时、全面的金融信息服务提供给客户, 不仅有助于提升服务人员的专业服务能力, 同时也提升了投资者服务质量, 为服务人员开展业务提供更有力的支持.

③ 专题资讯分析有助于提升机构用户的公司信

息管理水平.

本平台通过同业动态、重大事件跟踪等专题资讯分析模块, 能及时全面地为公司管理决策、机构监管分析等提供参考, 有效提升互联网时代下公司及监管部门等机构的信息管理水平, 防范市场风险, 提高管理决策能力.

综上所述, 本金融信息服务平台在资讯推送时效性上、内容质量上以及客户服务体验上均取得了不错的应用成效, 在大数据时代的金融信息服务创新起到良好的示范作用, 具有较大的行业推广价值.

4 结语

本文立足于证券行业的金融信息服务需求, 实现了一个集多源异构数据采集、分析、处理及发布全流程的金融信息服务平台. 该平台基于大数据建立了个性化的行业资讯推送、投资者情绪指数、股市预测等关键分析模型, 可为投资者提供更为精准、个性化的金融信息服务. 本文提出的平台架构及分析方法, 以证券投资信息服务为典型案例, 并取得较好的应用效果, 这种架构和分析方法还普遍适用于其他金融行业如银行、保险、基金等.

参考文献

1 Donovan S. Big data. Nature, 2008, 455(7209): 1–136.

2 程学旗,靳小龙,王元卓,等.大数据系统和分析技术综述.软件学报,2014,25(9):1889–1908.

3 Liao XW, Chen H, Wei JJ, et al. A weighted lexicon-based generative model for opinion retrieval. 2014 International Conference on Machine Learning and Cybernetics (ICMLC). IEEE. 2014, 2. 821–826.