

金融信息服务系统技术框架研究^①

俞 枫¹, 李洁华²

¹(国泰君安证券股份有限公司, 上海 200120)

²(上海市总工会职工援助服务中心, 上海 200072)

摘 要: 金融资讯数据是投资者进行投资决策、券商投研部门进行研究和撰写研究报告的重要依据。从金融资讯数据的数据整合与存储、信息检索和资讯推荐三个层次出发, 提出了一种创新的金融信息服务系统技术框架, 并在此基础上描述了国泰君安金融资讯与统一检索平台的构建方案。平台还在数据整合、数据存储、信息检索、去重机制、资讯推荐等多个方面实现了技术创新, 为金融行业构建信息服务系统提供了有益的示范与参考。

关键词: 技术框架; 数据整合和存储; 信息检索; 去重机制; 资讯推荐

Technical Framework of Financial Information Service System

YU Feng¹, LI Jie-Hua²

¹(Guotai Junan Securities Co. Ltd., Shanghai 200120, China)

²(Shanghai Municipal Trade Union Council Center of Aid and Service for Workers, Shanghai 200072, China)

Abstract: Financial information and data is not only an important reference for investors to make investment decisions, but also an important basis for the Securities Investment Research Department of research and research report writing. This paper from the three levels of data integration and storage, information retrieval and recommendation puts forward the technical framework of financial information service system for innovation. On this basis build the Guotai Junan financial information database and unified retrieval platform. The platform provides the demonstration and reference for the construction of information service system for the financial industry with the innovation of data consolidation, data storage, message retrieval, deletion of duplicated and information recommendation.

Key words: technical framework; data consolidation and storage; message retrieval; deletion of duplicated; information recommendation

1 引言

金融资讯数据是投资者进行投资决策、券商投研部门进行研究和撰写研究报告的重要依据, 为公司客户和投研部门提供及时、准确、易用的金融资讯数据一直是券商所面临的一项长期而艰巨的挑战。随着网络富信息化和大数据时代的到来, 目前金融资讯数据中已包含大量的结构化和非结构化信息, 且增量巨大, 特别在券商日益提升客户服务深度和广度的大背景下, 券商在金融信息服务方面所面临的挑战可以抽象为以下三个层面:

1)“存”: 即要运用数据整合技术, 实现对多种信

息系统、数据库、文档等异构金融资讯数据的无缝整合, 并保证数据存储的安全性、完整性和易用性;

2)“取”: 即要运用大数据处理的技术, 采用检索、去重等手段为用户提供全面、精确的信息检索服务;

3)“用”: 即要运用资讯推荐的技术, 通过建立用户数据使用模型, 进行个性化的信息数据推送, 实现以数据为核心的增值服务。

根据上述的业务抽象, 我们提出了一个包含三个层次的金融信息服务系统技术框架(如图 1 所示), 从低到高可划分为数据整合与存储层、信息检索层、资讯推荐和展示层, 这三个层面分别对应实现金融资讯数据

① 基金项目: 国家科技支撑计划(2012BAH13F03)

收稿时间: 2014-01-22; 收到修改稿时间: 2014-03-10

信息服务的“存”、“取”和“用”这三个层次的应用。



图1 金融信息服务系统技术框架

面对金融资讯数据的非结构化、传播迅速、自媒体丰富等特点, 国泰君安证券为了适应行业转型的需要, 于2012年初基于上述的金融信息服务系统技术框架开始构建国泰君安金融资讯和统一检索平台, 整合来自公司内部和互联网的各类结构化与非结构化金融资讯数据, 并采用清洗、转换、检索、去重、推荐等技术手段, 为公司客户、员工提供准确、及时、个性化的金融资讯数据服务。

2 数据整合与存储层

2.1 数据整合

该金融资讯及统一检索平台的构建首先就面临着对来自多种信息系统、数据库、文档等的异构数据进行无缝整合难题。XML技术作为数据整合的事实标准^[1], 具有与平台无关、易于扩展、交互性好和语义性强等特性, 因此, 我们选择了基于XML的数据集成模型来实现对各异构数据库的描述以及数据源间的数据转换, 并通过对数据的抽取、转换、清洗和装载来完成异构数据的整合工作。

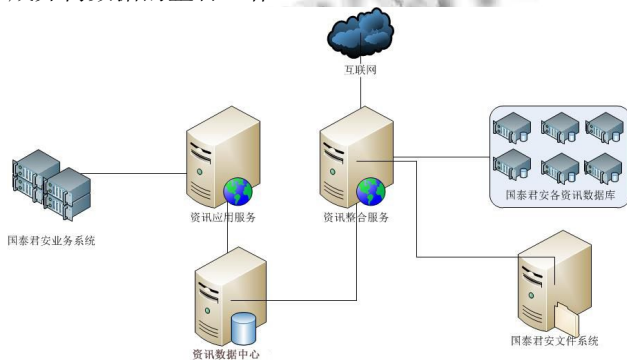


图2 国泰君安资讯及统一检索平台系统部署架构图

国泰君安金融资讯及统一检索平台的系统部署如

图2所示, 平台系统主要包括五大部分: 业务系统、资讯应用服务系统、资讯数据中心、资讯整合服务系统及各种分散的资讯数据库。其中, 业务系统通过调用资讯应用服务系统进行信息检索, 使用资讯信息进行推荐服务。资讯应用服务系统具有检索与索引功能, 负责对数据进行统一的索引与维护。资讯整合服务系统具备对结构化、非结构化数据进行整合的能力, 采用去重机制及资讯推荐技术实现资讯数据的分类汇总和规整, 并通过协议接口和前端页面展示为用户及业务系统提供信息服务。

其中, 资讯整合服务系统通过ETL技术整合分散在各数据库系统、文件系统以及互联网的金融资讯数据, 对数据进行采集和清洗, 并通过基于业务主体域的资讯整合策略将不同来源主体的资讯数据整合形成资讯数据中心。数据整合服务的主要过程包括: 首先将查询请求以XML Schema形式传递到数据抽取模块, 数据抽取模块将XML转换成SQL查询语句, 然后根据查询结果进行数据抽取, 最后把抽取出来的结果集转换成XML的格式传递给集成处理模块, 同样, 非结构化数据也需转化成XML格式, 再由集成处理模块对XML文档做集成处理并最终生成统一的资讯数据中心。

2.2 数据存储

平台作为金融行业信息服务系统, 数据量很大, 对数据的安全性要求也很高。Hadoop框架^[2]使用分布式文件系统(HDFS)作为低层存储支持, HDFS提供了一种高容错性和高吞吐量的海量数据存储解决方案, 其不停机动态扩容、数据自动检测和复制等特性^[3]为平台的大数据存取和数据的高安全性提供了解决途径。HDFS的文件分块存储特性使得在进行系统容量扩充时可依靠系统的分布算法自动实现数据块的迁移和容量的升级, 无需系统宕机或人工维护。HDFS所具有的数据自定义复制策略和数据一致性自动监测机制满足了数据的高安全性要求。HDFS资源最优分配和多副本访问机制大幅度提高了系统的数据读取速率^[4], HDFS对于单数据块的访问性能是传统存储方案的数倍。

该平台的HDFS数据存储模式如图3所示, 从上到下依次分为Daas、Paas、Saas三层。

(1)Daas(数据即服务层)主要负责数据存储和检索, 利用HDFS的灵活性、低延迟、分布性等特点将资

讯中心的数据经过规整后对外提供数据服务。

(2)PaaS(平台即服务层)主要负责数据和文件的访问及支持二次开发,统一认证由LDAP服务器完成,平台采用JDBC数据访问接口为业务系统屏蔽了异构DBMS的差异。

(3)SaaS(软件即服务层)采用用户层虚拟化技术实现了集中交易日志存储和分析系统以及历史行情数据管理和查询平台等对外提供多租户、可扩展的软件服务。

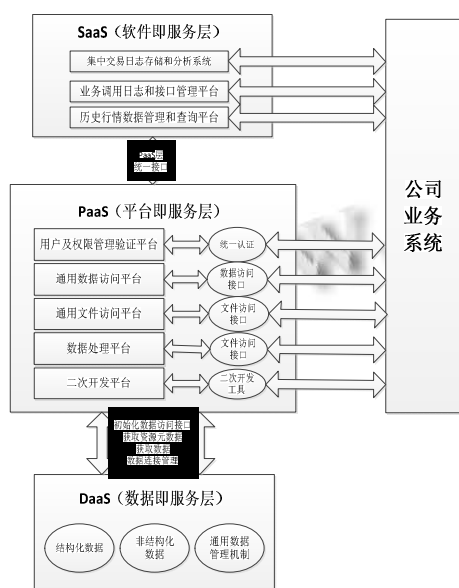


图3 平台 HDFS 数据存储模式

3 信息检索层

该资讯及统一检索平台整合后的金融资讯数据量很大,而且有许多资讯数据是由不同的信息披露主体、研究主体和行业媒体从不同角度对同一资讯所进行的加工和处理,因而平台面临着检索效率低、检索结果中存在大量的重复冗余信息的挑战。为了提高信息使用的效率和便利性,提高用户的体验度,平台运用大数据处理的技术,通过全文检索、资讯去重等手段为用户提供全面、精确的信息检索服务。

3.1 信息检索

目前,金融资讯数据中非结构化信息占信息总量的80%以上,而传统关系型数据库的字段检索技术对处理非结构化信息有先天不足,尤其是在对于海量非结构化信息的处理方面^[5]。由于全文检索数据库可支持重复字段、子字段以及变长字段,可实现对变长数

据、重复字段进行处理以及数据项的变长存储管理,在处理连续信息和非结构化信息时有着传统关系型数据库所无法比拟的优势^[6]。为此,平台运用全文检索技术来解决对非结构化信息的处理,基于开源的Lucene开发框架,通过对Lucene核心层与相关接口的定制开发来构建全文检索系统。

该检索系统以Lucene全文搜索引擎为核心,从功能上可分为索引、搜索与维护三部分。索引部分负责对数据库存储的数据进行处理,建立索引结构;搜索部分接收前端系统递交的检索请求,对索引进行查找;维护部分则负责对索引进行增加、修改、删除等维护工作。整个检索系统的实现过程包括:对文档进行预处理;进行分词并创建文档索引。对于中文分词, Lucene 采用的是二元分词法;提供搜索功能,即利用 Lucene 建立的索引进行搜索^[7]。

系统所采用的 Lucene 开发框架(如图4所示)包括 Lucene 核心功能模块与定制开发模块两部分。Lucene^[8]核心功能模块包括索引/检索器层、存储层以及倒排索引文件层,其中,倒排索引用来存储在全文搜索下某个词在一个文档或者一组文档中的存储位置的映射,是 Lucene 实现快速查询的核心技术。基于核心层之上的定制开发模块,包括词法分析层、文本解析层以及应用层。其中,文本解析层主要通过各种文档解析器对不同格式的文档进行解析,从而得到便于操作的文本文件;词法分析层则主要是将文本划分成词并选择合适的词来建立索引,用于中文检索时需要使用相应的中文分析器。

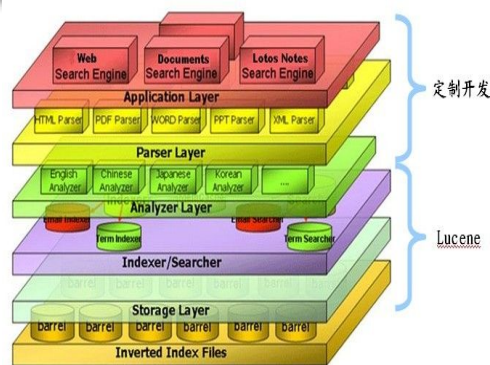


图4 Lucene 开发框架

3.2 资讯去重

为了获得更好的检索效果,系统还需要对每日入库超过3万条(篇)的各类金融资讯数据进行去重处理。

由于提高信息去重的效率对于提升检索系统的性能以及用户的使用体验意义重大,为此,针对金融资讯数据所存在的关键词集合小、重复率高以及文本结构特征相似的特点,我们采用了新的基于 SimHash 的去重框架,创新地提出了一种基于段落主题的文本去重机制(DRPKP 算法)^[9],使用文本数据的主题信息来比较其相似度,实现对于相同主题和雷同内容的金融资讯数据的归类,以进一步提高去重效果。

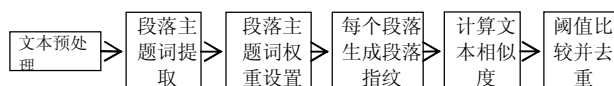


图 5 DRPKP 算法的去重流程图

SimHash 是局部敏感哈希(Locality Sensitive Hash, LSH)的一种,最早由 Charikar.M.S 于 2002 年提出^{[10][11]},是目前主流的文本去重算法,被包括 Google 在内的互联网公司广泛使用于开放领域的网页文本去重,具有计算速度快、存储空间小等优点。相比于 SimHash 算法,我们提出的 DRPKP 算法充分考虑了文本的结构以及特性的分布情况,SimHash 算法是一个文本产生一个指纹,而 DRPKP 算法是由文本中的每个段落产生一个指纹,因而一个文本可以表达为基于段落主题的指纹集合。对于同一个文本而言,基于段落主题的指纹集合比单一指纹包含更多的信息,这些信息在计算指纹集合的海明距离时可以把文本间的差异性放大,从而提高对文本相似度判断的准确率。DRPKP 算法的去重流程图如图 5 所示,该去重方法的整体步骤包括:根据文本的段落主题(Paragraph Key Phrase)提取每个段落的段落指纹,然后通过比较两个文本的段落指纹来计算两个文本的相似度,最后相似度超过设定阈值时则认为是重复文本,进行去重操作。

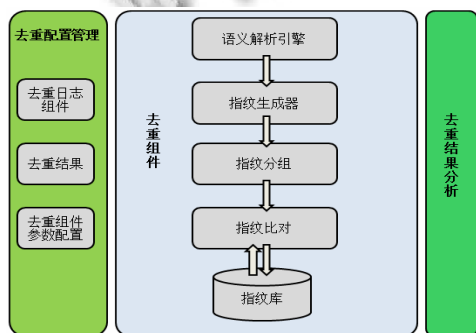


图 6 资讯去重框架

该资讯去重框架的整体结构如图 6 所示,主要包括:去重组件、去重配置管理、去重结果分析三部分。其中,去重组件通过语义解析引擎负责对资讯内容进行分词处理,指纹生成器基于 DRPKP 核心算法根据分词结果生成资讯的指纹特征。64 位的指纹则是按相同规则被均分 4 组进行索引存储。在指纹比对过程中,首先要根据抽屉原理进行数据计算的降维,并计算资讯指纹特征与指纹库的资讯指纹海明距离大于或等于 3 的比对结果。去重配置管理负责对资讯去重的结果进行日志记录,并可以检查资讯去重的结果。

DRPKP 算法是基于 SimHash 算法框架发展的,平台使用类 SimHash 算法框架较好地解决了系统的性能问题,同时通过信度、效度的质量控制方式来解决高性能去重的基础之上有效性问题。平台基于去重日志信息从信度和效度两个角度对当日的去重结果进行质量分析,并对去重效果进行评价。从国泰君安证券实际使用的资讯去重效果上来看,该去重机制带来的延迟不超过 3 分钟,重复资讯去除率超过 90%,去重的准确率和召回率都达到了预期效果,大大提高了用户的资讯使用效率和体验度。

4 资讯推荐和展示层

信息检索实质上是由用户主导的,呈现给用户的是一样的资讯和排序结果,无法主动针对不同用户的兴趣爱好提供相应的个性化服务,因而在信息爆炸的情况下,使得信息的利用率可能反而降低,容易出现信息超载的现象。为了进一步提升用户体验,该金融资讯及统一检索平台还建设了资讯推荐系统,采用基于用户行为聚类的推荐算法来实现个性化的资讯数据推送服务。这种个性化推荐是通过建立用户和资讯数据之间相互对应的二元关系,利用用户行为的相似性关系来挖掘每个用户潜在感兴趣的对象,进而进行个性化的推荐,其本质也是一种信息过滤。

该资讯推荐系统由用户信息的行为日志记录模块、用户偏好的模型分析模块和推荐算法模块三部分组成。其中行为日志记录模块负责在各业务接触点记录用户的各种行为信息,包括页面的停留时间、点击序列、内容浏览记录以及用户的个人信息和交易历史(来源于集中交易系统)、行情浏览历史(来源行情系统)等,这些信息是后续分析和资讯推荐的数据基础;用户偏好的模型分析模块负责对用户行为日志的分析,

对用户多角度的属性加以测算和评分,为每一个用户建立各自的多属性描述,并运用业务知识和数据挖掘工具对用户的众多属性评分进行聚类,也就是将具有相似行为模式的用户聚集在一起,该系统已按照用户的风险偏好、资产状况、持仓分布、交易活跃度、盈利能力、投资品种偏好、生命周期、资讯使用偏好、资讯使用历史等多个属性建立起了相应的分类分级用户数据使用模型,该模型的有效建立是整个推荐系统的难点;推荐算法模块则是利用组合算法从资讯中心中根据分类分级用户模型为用户实时计算出各资讯的客户兴趣度,并返回其中前 N 条给业务前台进行集中展示,推荐算法模块是整个推荐系统的核心环节。

目前,系统所采用的是一种组合推荐的算法,其中针对用户过去投资行为记录和对某板块股票的偏好(如持仓分布信息、历史交易记录),对相应板块的金融资讯数据进行基于内容的优先推荐;针对用户的投资偏好、资讯使用偏好(如小市值股票偏好、区域板块股票偏好、研究报告偏好、数据分析偏好)等规则向特定用户进行基于规则的推荐;根据用户所在的分类分级用户群的资讯使用偏好和习惯对其中的个体进行协同过滤推荐。这种组合推荐算法可利用评分模型对三种推荐的结果进行组合和综合评分,可以避免单一推荐方法的局限性,推荐的效果更佳。

国泰君安证券于 2012 年开始以“君弘”品牌为载体实施客户的分类分级服务,即根据客户不同类型的需求对客户实施分类的精细化管理,对在同一客户类型中针对客户的不同价值贡献度匹配相应的公司分级服务资源。作为落实公司客户分级服务和分类管理的重要举措,国泰君安资讯及统一检索平台推出了资讯推荐服务,通过建立分类分级的用户数据使用模型实现了个性化的金融资讯数据推送,为客户提供以数据为核心的增值服务,极大地提升了客户的满意度。

5 结语

国泰君安资讯及统一检索平台基于创新的金融信息服务系统技术框架,按照数据整合与存储、信息检索、资讯推荐和展示三个层次进行构建。该平台使用 XML 技术实现了对多种异构金融资讯数据的有效整合,采用 HDFS 文件系统实现了高容错性和高吞吐量的海量数据存储,通过建立全文索引及采用创新的基于段落主题的文本去重机制(DRPKP)提高了信息检索

的效果。平台于 2013 年 5 月正式上线,已整合了公司内部生产、外部采购、和来自互联网的各类结构化与非结构化的各类金融资讯数据,整个数据量已经超过 300G,为投资者和投研部门提供了全面、精确的信息资讯,并基于用户行为聚类(分类分级用户数据使用模型)的组合推荐算法实现了个性化的信息推送,提升了金融信息服务的智能化和个性化水平。该平台为用户所提供的以数据为核心的增值服务,大大提升了金融资讯数据的使用效率和使用效果,成为提升券商的投研水平和客户服务效率的新的手段。该平台所基于的金融信息服务系统技术框架,结构清晰,可行性强,可以在金融行业中进行广泛使用,平台还在数据整合、数据存储、信息检索、去重机制、资讯推荐等多个方面实现了技术创新,对金融行业信息服务系统的构建提供了有益的示范和参考。

参考文献

- 1 赵玉玲.基于 XML 的数据集成技术的研究与实现[博士学位论文].长春:吉林大学,2009.
- 2 Apache. Hadoop. <http://hadoop.apache.org/>. [2013-06-18].
- 3 黄晓云.基于 HDFS 的云存储服务系统研究[博士学位论文].大连:大连海事大学,2010.
- 4 田秀霞,周耀君,毕忠勤,彭源.基于 Hadoop 架构的分布式计算和存储技术及其应用.上海电力学院学报,2011,27(1):70-74.
- 5 郎小伟,王申康.基于 Lucene 的全文检索系统研究与开发.计算机工程,2006,32(4):94-96,99.
- 6 张校乾,金玉玲,侯丽波.一种基于 Lucene 搜索引擎的全文数据库的研究与实现.现代图书情报技术,2005,21(2):40-43,48.
- 7 吴鹏飞,马凤娟,李文革,郭鹏.开源全文搜索引擎 Lucene 本地化实践研究.现代图书情报技术,2009(4):19-22.
- 8 Apache Lucene. <http://lucene.apache.org/>. [2013-02-22].
- 9 俞枫,王引娜.基于 DRPKP 算法的文本去重研究与应用.微型电脑应用,2014,30(1):58-60.
- 10 Charikar MS. Similarity estimation techniques from rounding algorithms. Proc. of the Thirty-fourth Annual ACM Symposium on Theory of Computing. New York, ACM. 2002. 380-388.
- 11 张校乾.基于 Lucene 的全文检索系统的研究与应用[博士学位论文].大连:大连理工大学,2005.