计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200900083

面向大数据分析的智能交互向导系统

余乐章[1,2]　夏天宇[1,2]　荆一楠[1,2]　何震瀛[1,2]　王晓阳[1,3]
1 复旦大学计算机科学技术学院　上海 201203
2 上海市数据科学重点实验室(复旦大学)　上海 200433
3 上海智能电子与系统研究院　上海 201203
(19212010048@fudan.edu.cn)

摘　要　传统的大数据工具一般为专业数据分析人员打造,具有难以上手、操作交互性差以及不够智能化等特点。而智能交互向导系统是针对大数据交互式分析系统目前存在的问题而研制的一套大数据分析辅助工具。系统既研发了用户意图理解、数据抽样及列推荐、可视化推荐、分析方法推荐等核心关键技术,也拥有良好的图形化界面与人性化的智能交互体验。在满足用户多种交互式分析需求的同时,还具有极高的响应速度。不仅可以随时回溯到分析流程任意一步重新选择方法的执行流程,还可以通过接口与各种分析应用快速集成以部署应用于不同场景。经过实验测试,系统的平均交互时间均在3s以内,且与传统分析方法相比系统交互的执行时效加快了3倍左右。通过用户用例测试,系统的满意度相比传统工具更加优秀。智能交互向导系统通过在易用性、时效性、可交互性和智能性等方面的探索,让不同基础的用户群体都可以使用此系统完成所需的大数据分析目标。

关键词:大数据系统;智能交互;数据分析;方法推荐;用户意图
中图法分类号　TP311.5

Smart Interactive Guide System for Big Data Analytics

YU Yue-zhang[1,2],XIA Tian-yu[1,2],JING Yi-nan[1,2],HE Zhen-ying[1,2] and WANG Xiao-yang[1,3]
1 School of Computer Science and Technology,Fudan University,Shanghai 201203,China
2 Shanghai Key Laboratory of Data Science,Fudan University,Shanghai 200433,China
3 Shanghai Institute of Intelligent Electronics and Systems,Shanghai 201203,China

Abstract　Traditional big data tools are generally built for professional data analysts,and they have the characteristics of being difficult to get started,poor operation interaction,and not intelligent enough. The intelligent interactive guidance system is a set of big data analysis auxiliary tools developed around the current problems of the big data interactive analysis system. The system not only develops core key technologies such as user intention understanding,data sampling and column recommendation,visualization recommendation,and analysis method recommendation,but also has a good graphical interface and a humanized intelligent interactive experience. While meeting the user's multiple interactive analysis needs,it also has a very high response speed. Not only can you go back to any step of the analysis process to reselect the method execution process at any time,but you can also quickly integrate with various analysis applications through the interface to deploy and apply to different scenarios. After experimental tests,the average interaction time of the system is within 3 seconds,and the execution time of the system interaction is accelerated by about 3 times compared with the traditional analysis method. After using case testing,the system is also more satisfying than the use of traditional tools. Through the exploration of ease of use,timeliness,interactivity,and intelligence,the smart interactive guide system allows users of different basic groups to use the system to complete the required big data analysis goals.

Keywords　Big data system,Smart interaction,Data analysis,Method recommendation,User intention

1　引言

当今社会,数据量与日俱增,数据储量的规模一般在TB级上下。在生产生活中,用于处理大数据的工具逐渐被越来越多的公司或个人所需要,尤其是大数据分析工具对于企业和商家来说,成了查漏补缺、拉高业绩的重要引擎。

而目前市场上已有的传统大数据分析工具,无论是OLAP系统还是BI工具,虽然各有优点,但也都存在着各种

到稿日期:2020-09-10　返修日期:2021-01-29　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家重点研发计划资助项目(2018YFB1004404)
This work was supported by the National Key R&D Program Funded Project of China(2018YFB1004404).
通信作者:荆一楠(jingyn@fudan.edu.cn)

问题。有些工具缺乏时效性,往往需要大量的时间进行数据计算;有些工具不够智能,只能完成单一的特定分析动作;还有些工具没有可视化的图形界面,对于数据的展现不够直观。

最重要的是,这些工具都过于专业化,只适用于行业中拥有大量领域经验的专业数据分析工程师,而普通用户只能对着想要处理的大量数据望洋兴叹。如何使大数据领域的专业技术以用户需求为导向,让用户理解并操作完成其大数据处理目标成了当下需要解决的问题。

为了满足市场上日益增长的需求,本文提出了一种面向大数据交互式分析的辅助工具——智能交互导向系统。该系统具有非常强的易用性、可交互性、智能性与时效性。通过对该系统及关键技术的研发,实现对用户分析意图的理解,在数据、分析方法模型统一建模的基础上,根据用户意图实现数据、可视化图表、分析方法等的智能推荐,从而给用户提供易用、流畅的大数据分析体验,以提高分析效率,为各行业、各层次基础的有大数据分析需求的人员提供更简便高效的大数据分析工具。

智能交互向导系统可独立于其他系统之外,有着良好的高内聚性与低耦合性,方便通过接口集成并可以利用容器进行部署。同时,该系统还具有良好的延展性,可以融合更多的工具对系统自身进行扩展。

本文的贡献主要包括以下几个方面:

(1)设计了一套大数据分析辅助工具,其易用性和交互性可以满足任何基础用户的数据处理需求。

(2)提出了一种系统分析流程,可以分析理解用户意图,找到用户最感兴趣的数据范围并推荐不同的分析处理方法。

(3)集成了列推荐、可视化推荐、分析方法推荐等关键技术,解决了一般大数据分析工具机械化的问题。

(4)与传统分析方法进行比较,同时向不同层次的用户进行了满意度调查,以此展示智能交互向导系统的时效性与实用性。

本文第 2 节介绍了目前大数据分析工具的现状及存在的问题;第 3 节介绍了系统的设计思想;第 4 节着重讲述了系统和关键技术的实现;第 5 节通过时效实验与满意度调查展示了系统的优势;最后总结全文。

2 相关研究

2.1 OLAP 系统

OLAP 系统[1]在生产环境中的研究应用主要集中于支持用户分析决策,具体表现为企业的大数据分析和统计应用,如年度读书报告、年度销售报告和年度最受欢迎手机品牌 Top5 等。分析决策人员根据 OLAP 系统得出的统计结果,分析企业经营情况或者制定相应销售计划等。

OLAP 系统支持的查询一般涉及大量的数据,这些查询也往往包含了聚合、排序、JOIN 等复杂操作。因此,查询吞吐量和响应时间是关键性能指标。许多 OLAP 系统结合各种技术很好地满足了这些指标。

S-OLAP 是一个用于分析序列数据的 OLAP 系统,它提出了与传统 OLAP 系统不同的序列数据立方体(sequence data cube)[2-3],还认为一个序列不仅可以通过其构成项目的属性值来表征,还可以通过其拥有的子序列/字符串模式来表征。S-OLAP 也提供了方便用户使用的交互界面,能够使结果实时响应更新,从而使用户可以对序列数据进行探索性分析。

L-Store 是一个实时的 OLTP 和 OLAP 系统,它通过引入基于世袭的数据存储结构(lineage-based),在单个处理引擎中结合了对事务性和分析性工作负载的实时处理[4]。实验证明,与当时的最新技术相比,其具有优越性。

然而,现有 OLAP 系统未能实现分析过程的自动化,仍然需要用户手动进行查询,而普通用户在面对海量的数据时不知道应该从何处入手,这样就导致了使用 OLAP 系统的门槛较高。

2.2 BI 工具

BI 工具[5-6]是一类由数据仓库、查询报表、数据分析等模块组成的用于帮助数据分析专家进行企业决策的分析工具,目前市场上存在许多功能强大的 BI 工具,主要有以下几种。

(1)Tableau[1),一个可视化图表美观且操作简单的 BI 工具。Tableau 可以连接数据库,或者导入 Excel 文件数据。用户通过拖拽数据的维度和度量到工作区,来形成可视化图表,可以改变图表的外观。这些操作都通过鼠标完成,十分方便简单。但是,对于分组、筛选、下钻等分析操作,需要用户编辑公式。

(2)PowerBI[2),一款由微软开发的 BI 工具。PowerBI 支持各类数据源,除了 Excel 和 CSV 文件,它还支持 Access、SQL 数据库、HDFS、第三方 API 等。如果数据表之间需要建立联系,则需要在 Excel 中编写函数来实现,而 PowerBI 只需要通过拖拽来完成,非常便捷。

(3)FineBI[3),与 Tableau 类似,支持通过拖拽维度和度量来完成可视化图表的创建,支持进行数据抽取和数据索引建模,可显著提高数据的计算速度,可实现离线查询。

(4)Superset[4),Airbnb 开源的大数据可视化平台,目前由 Apache 孵化。后端几乎支持所有主流的数据库,支持复杂的权限控制和管理(OpenId,OAuth 等),满足企业内部不同部门人员对数据的权限要求。数据的展示完全可控,可自定义展示字段、聚合数据、数据源等。

(5)Metabase[5),一个轻量级的开源 BI 工具。易于部署,支持邮件通知。最重要的是,用户即使不熟悉 SQL,也能够通过提出的问题对数据进行探索,降低了使用门槛。相对地,其功能不如 Tableau 这种大型的 BI 工具强大。

2.3 研究现状

除了以上提到的系统和工具,业界还存在许多大数据系统。如 HaoLap,一个基于 Hadoop 的 OLAP 系统,采用多维模型来映射维度和度量,通过分区和线性化算法来存储维度和度量,并用 MapReduce 执行 OLAP 操作[7]。HaoLap 在不同数据集大小和不同复杂程度的查询上都有不错的表现。又

1) https://www.tableau.com/

2) https://powerbi.microsoft.com/

3) https://www.finebi.com/

4) https://superset.incubator.apache.org/

5) https://www.metabase.com/

如 BDA 大数据分析系统[8]，通过分析大数据技术发展现状[9]，提出了一个可以提供通用智能的大数据分析算法库、高可复用的分布式计算框架，以及面向数据流图的新颖交互模式的高效大数据分析引擎。该系统给用户提供了简单易用的使用模式和交互方法，但这里的用户仍然是指掌握了大数据技术和知识的工作人员，对于普通用户，其易用性难以保证。

上文提到的 OLAP 系统，虽然有着很好的时效性，但是使用门槛较高，需要用户具备一定的数据科学和数据分析专业知识。而另一种 BI 工具，虽然可以使不熟悉 SQL 的用户能够简便地创建可视化图表进行数据探索分析，但是这些分析过程涉及大量的试错，是一个耗时耗力的过程。而智能交互向导系统相比这些系统，不仅实现了在大数据平台上的秒级 OLAP 查询，并还聚焦于用户交互的智能化。用户不需要具备专业的数据分析知识，只需要在 Web 界面中选取感兴趣的维度，系统将会自动完成 OLAP 查询，并利用可视化推荐技术展示查询的可视化结果。如果用户想分析某一个查询结果或者依据该结果进行决策，只需要点击相应的可视化图表，则智能交互向导系统又会推荐可以进行分析的方法，并返回分析的结果供用户参考。由此，智能交互向导系统可以在交互式分析过程中为用户提供易用、智能的引导。

3 系统设计

3.1 设计思想

如图 1 所示，系统的主要设计思想是，作为用户与大数据分析计算工具的中间辅助系统，智能交互向导系统既包含与用户完成交互的前端界面，又通过 API 中间层完成与下层 Spark，Hive 等大数据处理工具的集成调用。

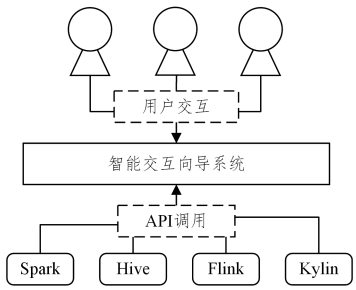

图 1 系统设计图

Fig. 1 System design drawing

整个系统可以为用户实现一套完整的、满足处理需求的大数据分析流程。用户只需要操作前端简易的可视化界面即可完成后台的数据提取、处理、反馈等一系列复杂的数据分析动作。

3.2 工作流程

典型的交互式分析流程一般包括几个过程：收集转换数据、可视化数据、创建模型、诠释结果等。如图 2 所示，智能交互向导系统也符合经典的流程模型，其主要关注用户兴趣数据和方法选择以及可视化图表推荐等步骤，通过列推荐、方法推荐等方法优化分析流程，以得到最优的分析结果。

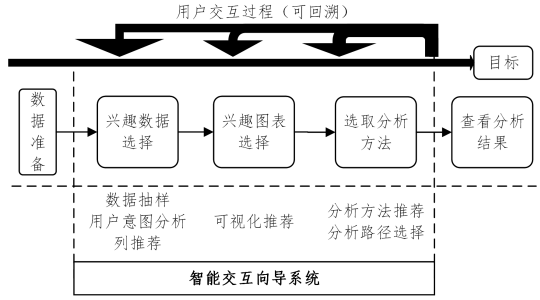

图 2 智能交互向导系统数据分析流程图

Fig. 2 Smart system data analysis flow chart

智能交互向导系统会以用户的业务数据为源数据进行准备处理，通过数据抽样的方式，快速读取指定范围的数据样本作为分析样本。通过用户意图分析与列推荐的方法为用户智能推荐他可能感兴趣的数据列与范围。在用户完成兴趣数据选择后，根据用户选择的数据范围，自动进行可视化推荐，推荐最适合展示数据维度与变化的图表供用户选择，以进行下一步分析。收到推荐的可视化图表后，系统的分析方法推荐部分会根据已有的分析方法知识库推荐出用户想要的几种分析方法以供用户选择。分析的结果也会以图表的形式展现给用户，用户如果对结果不够满意，也可以随时回溯到分析流程的任意步骤，重新进行流程分析。

3.3 流程架构

整套系统设计的流程架构如图 3 所示，系统以分布式大数据平台 Hadoop 为基础，以 Hive 等为数据存储的仓库，并且辅以 Kylin 等计算框架为引擎对大数据进行分析处理。同时，在逻辑部分，前后端服务框架相辅相成，组成智能交互向导应用，根据用户的操作来实时分析用户的行为意图，并进行理解与预测，将用户最想要的选项或是结果呈现给用户以供用户交互选择。

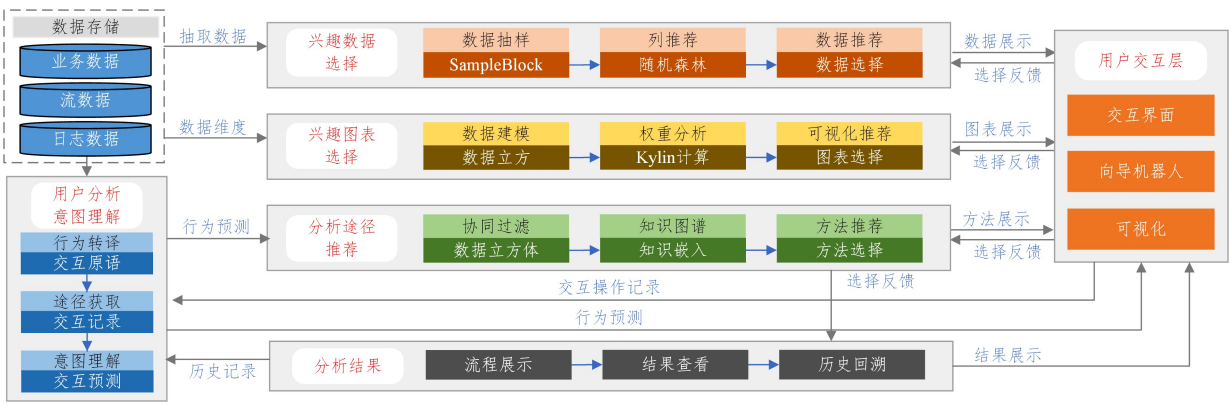

图 3 系统流程架构图

Fig. 3 System flow architecture

因为智能交互向导系统具有大数据分析的完整生态，所以低耦合性和高内聚性的特点可以让这套系统独立于其他平台系统之外，自由地嵌入到任何需要的示范应用之中。并且系统连接时所需的接口也与一般标准相同，兼容性非常好。同时系统还具有极高的延展性，可以融合更多大数据分析的方法和工具，或是与其他优秀的存储系统、计算系统相整合，配合实际应用进行更好的服务。

以面向电商商家的智能运营辅助为例，整套智能交互向导系统以智能向导机器人前端交互插件的形式嵌入到网页之中。点击交互机器人便可以跳出交互式分析向导窗口，并且插件和窗口与原有网站平台相互独立，互不影响，如图 4 所示。

图 4 智能向导机器人

Fig. 4 Smart guide robot

4 系统和关键技术的实现

下文将以一个面向电商商家的智能辅助运营分析应用为例，来介绍智能交互向导系统的实现方法以及如何帮助用户完成大数据分析流程。

4.1 数据准备

4.1.1 数据存储

在执行分析流程之前，需要对指定系统用户的数据做好准备工作。系统的业务数据主要分为批处理数据和流数据两种格式。智能交互系统会提前将批处理数据存储在 HDFS 文件系统中，以供后续分析计算，并将业务数据映射到 Hive 数据表中存储，当需要使用时由系统后端通过 Spark 或 Kylin 等工具从 Hive 中取出进行使用。而流数据则通过 Flink 流数据计算引擎进行读取，并实时对流数据进行具体的计算分析。除了业务数据外，系统还利用 MySQL 作为元数据的存储数据库，并将涉及到的元数据信息设计成不同的关系型数据表，通过相应接口进行访问。日志数据也都被定时保存在 MySQL 中。

以电商系统业务数据为例，批处理数据就是顾客购买商品记录的关系型数据表，也可以理解为卖家的销售记录。而对于电商系统每时每刻新增的数据，如新增购买记录、用户评论等这类流数据，系统会按照之前提到的流程存储并通过 Flink 分析。

4.1.2 样本抽样

当用户打开系统前端时，智能交互向导系统会根据用户 ID 从 Hive 表中准备好用户所拥有的业务数据，在商家应用中即该卖家的商品销售记录等。因为数据的规模庞大，为了解决时效性问题，系统基于 Verdict[10] 中间件进行了改造，提出了一种基于抽样块（Sampling Block）的面向交互式分析抽样方法。该方法的主要思想是基于用户查询负载，将大数据集分割成小的 Sampling Block，生成更精细的样本，避免运行时读取无关样本点，提升了查询准确率和响应速度[11]。在系统的数据准备阶段，后端便自动调用抽样块管理模块进行离线样本构建，通过将 Hive 中存储的用户数据集合抽样成小的 Sampling Block，再存入 Hive 中，用于接下来的处理分析工作。

4.2 兴趣数据选择

4.2.1 列推荐方法

当系统在后台自动为用户提取出样本数据后，会智能地为用户推荐他可能感兴趣的数据列和范围，这里就用到了列推荐的技术。

列推荐算法的实质就是给定由大量数据组成的数据集 D，采用随机森林特征选择的方法，根据系统给定的一些关键数据计算数据列的重要度。这些关键数据会根据用户的历史记录、意图分析和自主选择等因素智能生成。在之后对计算出的数据列的重要性进行 Top-K 排序，将分析过滤结果中最重要的两列数据返回给用户。再针对这两列数据进行有效的数据范围分析过滤，最后将结果自动可视化展示在前端，展示系统认为用户感兴趣的数据列及数据范围。用户此时也可以根据自己的想法与系统进行交互，在系统的数据展示界面进行拖动、点击、选取。如图 5 所示，在电商商家系统的应用中，可以选择二月份到五月份的店铺商品销售量情况，来对这部分数据进行进一步的分析操作。

图 5 数据列及范围推荐

Fig. 5 Data column and range recommendation

4.2.2 用户意图理解

针对用户分析意图理解的问题，智能交互向导系统使用了一套独有的分析模块。在用户进入前端页面后，用户的每一步交互操作都会作为交互历史被系统记录在后台，并通过日志形式进行存储。同时，系统会根据建立好的已有的用户意图理解模型来实时分析用户的意图表达，并基于此提出预测的分析流程方法。用户意图理解模型是建立在人为设计的基础上，通过规则化的定义对用户行为作出解释。以电商系统场景举例，页面销量区域处的用户互动一般与系统分析方法中的销售量预测方法关联度高，而店铺支付金额区域的互动又一般与店铺收入检测的分析方法有关。

举例来说，如果用户的鼠标在前端页面的销量情况部分

停留时间较长，通过系统的用户意图理解模块分析，小机器人就会自动为用户推荐销量预测的分析方法并推荐选择销售量类型的数据范围。推荐的方法会实时显示在小机器人下方，如图 5 所示。通过用户意图的理解和预测，可以帮助缩短分析所耗费的时间，同时帮助实现更精准的数据预取，以进一步提升分析效率。

4.3 兴趣图表选择

可视化在数据分析过程中是一种非常直观的数据观察和分析手段，但数据的维度往往十分影响图表的合适度与用户体验。智能交互向导系统采用基于规则的方法[12]，根据不同数据类型和图表类型定义一个可视化图的权重[13]，然后根据这个权重构建由可视化图结点组成的偏序图[14]。在构建出各个可视化图结点的偏序关系后，可以计算出每个结点的得分，若结点没有出度，则其得分为零，否则其得分计算方式为：以该节点为起点，以有向边组成的路径上所有结点的得分加上该路径上所有有向边的权重之和。例如，假设存在两个有相同含义横纵坐标的可视化图 a 和 b，如果 a 的权重大于 b 的权重，则在偏序图中表示为 a 结点引出一条有向边指向 b 结点。最后根据不同图表的得分选择 Top-K 的图表推荐给用户，图 6 为推荐给用户不同数据标签所生成的图表。

图 6 兴趣图表推荐

Fig. 6 Interest chart recommendation

为了提升可视化图表推荐效率，系统会基于 Kylin 计算平台对 Hive 中原有的大数据的不同维度预先进行数据立方(Cube)操作，并将聚合的结果数据存储在 HBase 数据库中。在智能交互向导系统收到用户选择的数据列及范围后，便会通过指定接口将信息传递到 Kylin。Kylin 会根据系统设定的规则进行查询读取，快速生成最适合分析展示这部分数据的图表，图表类型包括柱状图、扇形图和折线图等。用户可以选择自己最感兴趣的可视化图，进入分析流程的下一个环节。

4.4 分析方法推荐

在收到用户的可视化图后，系统的分析方法推荐部分便会进行自动的方法推荐。根据用户选择的不同可视化图与之前用户意图分析的结果，智能交互向导系统会推荐相应可行的分析任务方法。如在文中所指出的卖家系统应用中，如果用户选择了销售量总和与性别作为标签的可视化图，那么系统会智能推荐销量预测或是检测异常值的分析任务等供用户选择。

对于这些任务采用的方法，智能交互系统采用了一种基于协同过滤和知识嵌入的推荐方法，即构建数据集和模型的交互矩阵。每个分析任务都是在某个数据集(或子数据集)上采用某种机器学习模型进行的，因此可以计算各种模型在同一个数据集上的得分，从而构建一个数据集-模型的交互矩阵。矩阵中的元素值被用来衡量模型对于数据集的合适程度，举个例子，如果模型 m 在数据集 i 上的得分很高，就说明模型 m 适用于数据集 i，使用模型 m 可以在数据集 i 上得到很好的分析结果。矩阵中的每一行都是一个数据集在不同模型下执行某个分析任务时的得分，从而可以挑选 Top-K 得分的分析方法作为推荐结果。

此外，为了提高分析方法特征向量之间的区分度，系统手工构建了关于分析方法的知识图谱，通过知识嵌入的方法将图谱转换为向量[15]，并将其叠加在经过矩阵分解后的交互矩阵上以形成增强矩阵，这样可以有效提高分析方法特征向量之间的区分度，解决分析记录的水平参差不齐等问题，并最终提升了得到的推荐结果质量[16]。系统支持的分析方法主要分为 3 个方面，分别为回归任务、聚类任务和异常检测任务的分析方法，固定为处理这几种任务的常见机器学习模型，如回归任务的线性回归(Linear Regression，LR)、支持向量机(Support Vector Machine，SVM)、聚类任务的 K 均值(K-means)，以及异常检测的主成分分析(Principal Component Analysis，PCA)、孤立森林(Isolation Forest，IF)等。后续也可以对分析方法及对应算法进行增量的更新以满足多种分析需求。

通过方法推荐模块的分析，由用户选择的数据、图表将转化成系统所预测出的用户最可能需要的 3 种分析方法展示在前端并供其选择。如可以选择对商品未来销售量进行预测或者对过去销售量的阈值情况进行异常检测等，推荐结果如图 7 所示。

图 7 分析方法推荐

Fig. 7 Analysis methods recommendation

4.5 分析结果与历史回溯

智能交互向导系统会在用户选择完想要的分析方法后，给用户展示系统将要执行的依据此方法制定的分析流程。待用户确认之后，经过计算，以图表形式将分析结果展示给用户。图 8 给出了此次用户选择的销售量预测分析方法的结果。

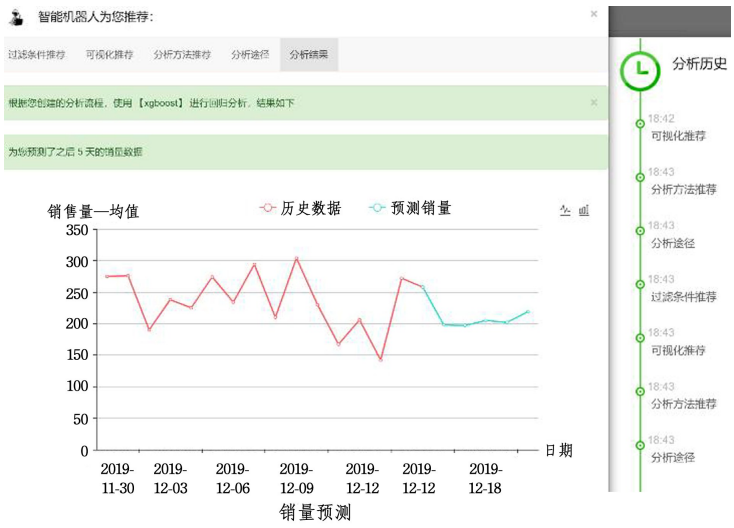

图 8 分析结果展示

Fig. 8 Analysis result display

如果用户对分析结果不够满意，也可以随时退回到之前分析流程的任何一个步骤。不同于其他大数据工具，智能交互向导系统简易的接口信息可以方便地将大数据分析步骤中每一步的数据信息都保存下来，用户的操作历史记录也始终被存储在数据库中以供使用。当用户处于分析流程的任意环节时，都可以查看已经执行的分析流程并随时可以回溯到任意步骤。如图 8 中的电商应用所示，随时可以点击右侧的分析历史更改自己的分析需求，重新选择想要的数据或者分析方法。

5 实验对比

5.1 实验设置

实验对比对象为以下两个系统。

(1) Hive[17]：基于 Hadoop 构建的一个数据仓库系统，提供 SQL 查询存储在 HDFS 的数据，可以将结构化的数据文件映射为数据库表。

(2) 智能交互向导系统(简称 Smart)：建立数据、分析方法与途径的推荐模型，设计了数据采样、数据立方等时效优化机制，同时具有高时效性。

实验的硬件环境是由 1 个 master 和 4 个 slave 组成的服务器集群，他们的配置都是相同的。系统是 Ubuntu 5.4.0-6ubuntu1～16.04.12，CPU 是 Intel(R) Xeon(R) CPU E5-2690 v2@ 3.00GHz，内存为 32GB。

实验使用的数据是由淘宝购物网站家居商品数据和人工模拟生成的购买记录合成的数据表。数据表包含 1000 万行(约 100GB)，每行有 8 个维度和 3 个度量，维度分别为商品名、商品类型、出售日期、购买用户、用户性别、用户所在省份、用户民族以及商品描述，度量分别为商品价格、销售量、商品折扣。具体如表 1 所列。

表 1 数据表示例

Table 1 Dataset example

item name	sold date	…	price	quantity
refrigerator	2019-08-21	…	3490.0	1
storage box	2019-05-05	…	268.0	4
…	…	…	…	…

实验在有向导支持下(在智能交互向导系统支持下)和无向导支持下(在比较对象 Hive 系统支持下)，对于同样的交互式分析用例，分别对比执行时效以及系统的交互时间。这两个性能指标的含义如下。

执行时效具体指在达到同样的分析效果下交互式分析过程所耗费的总时间，Smart 系统中包括了用户考虑的时间，而 Hive 中包括了专业人员构思查询语句的时间。交互时间具体指在以上数据规模下，智能交互向导系统的平均交互响应时间，这里只是计算系统在接受用户输入后响应的时间，而不包括用户在分析过程的不同阶段的思考时间。

实验使用了两个交互式分析用例(Session)。Session1 为商品销量异常检测，即通过数据筛选和可视化推荐后，选择商品销量异常检测的分析途径，最后获得分析结果的交互过程。Session2 为性别分析及销量预测，即通过数据筛选和可视化推荐后，选择性别分析和销量预测的分析途径，最后获得分析结果的交互过程。

实验对象在两个 Session 中的执行时效实验结果如图 9 所示。Hive 和智能交互向导系统在 Session1 下的执行时效分别为 214.90s 和 58.75s，在 Session2 下的执行时效分别为 178.56s 和 50.94s，可以看出智能交互系统的执行时效明显高于 Hive 的执行时效。智能交互向导系统在两个 Session 下的平均交互时间如图 10 所示。系统分析流程存在多个阶段，故取所有阶段的平均值。在 Session1 和 Session2 下的平均交互时间分别为 2.84s 和 2.62s。

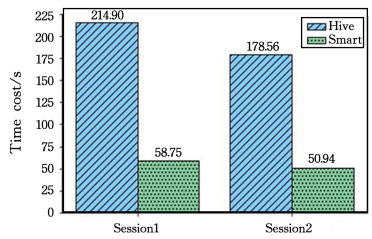

图 9 执行时效

Fig. 9 Execution time

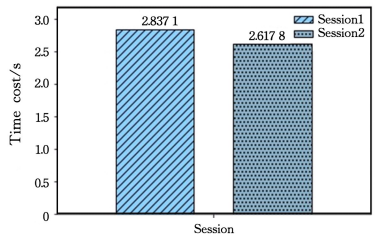

图 10 平均交互时间

Fig. 10 Average interaction response time

5.2 用户满意度的调查

除了与 Hive 进行了对比实验，本文还进行了用户满意度调查。参与者将使用该系统进行一系列 Session 的分析过程，最后根据使用情况对系统进行评分。实验者邀请了 100 位在数据分析领域有着不同知识掌握程度的学生作为参与者。他们中有半数以上的专业是与数据库技术相关的，其余参与者专业与数据库技术无关。专业相关的参与者大部分使用过数据库，掌握某种编程语言，半数使用过 BI 工具。而专业无关的参与者大部分没有使用过 BI 工具。受调查的参与者的知识层次各不相同，使得调查具有一定的普适性。

参与者被要求在同一数据集下使用 Tableau 和 Smart(智能交互向导系统)完成分析流程，并根据分析过程和结果进行满意度评价。实验参与者的满意程度用 1 到 4 的分数来衡量，从低到高依次为不满意、一般、满意、很满意。

同样，实验选取了 3 个 Session。Session1 为商品销量异常检测；Session2 为性别分析及销量预测；Session3 为根据顾客进行聚类。

用户满意度的得分情况如表 2 所列。实验参与者表示尽管 Tableau 提供了更美观的可视化图和方便的变换操作，但是分析操作难以上手；而智能交互系统可以提供一站式的分析流程，而且往往可以得到令他们满意的分析结果。然而，也有实验者提出系统现阶段支持的分析方法和分析路径太少，这也是系统未来的改进方向。

表 2 用户满意度

Table 2 User satisfaction

	Session1	Session2	Session3
Smart	3.49	3.34	3.16
Tableau	2.84	2.72	3.09

5.3 实验结论

智能交互向导系统在执行时效上较 Hive 提升了 3 倍以上，平均交互时间控制在 3s 以下，这说明系统拥有良好的时效性和交互性。系统在用户满意度的调查中的表现也优于 Tableau，证实系统拥有良好的易用性和用户友好性。

结束语 本文提出了一种面向大数据分析的智能交互向导系统，其优点在于：在易用性方面，系统集成便利，并且分析流程全程以可视化的图形界面进行操作，简单便捷，适合任何基础的用户使用；在交互性方面，用户意图理解模块可以理解用户的每一步操作，并作为系统推荐方法结果的依据；用户在分析流程中也可以随时回溯到之前的步骤重新进行操作；在智能性方面，借助于列推荐、可视化推荐、方法推荐等一系列智能算法，系统 可以更好地贴全用户的意图，从而对大数据进行分析处理；在时效性方面，通过数据抽样、离线保存配置数据以及算法的改进等，整个系统的响应时间达到了秒级。通过与同类型数据处理工具实验进行比较，系统的响应时间也非常优秀。经调研，系统也得到了不同专业基础的用户的认可。

当然，系统也存在一些问题，如系统目前的交互时间可以进一步减少，方法推荐所支持的方法数量不够多，推荐的算法模型精确程度也有提高的空间。在下一步的工作中，系统将会在不同的应用场景下进行部署，增加功能性并提升系统的整体性能，以针对不同的需求更好地完成大数据分析流程。

参 考 文 献

[1] CHAUDHURI S, DAYAL U. An overview of data warehousing and OLAP technology[J]. ACM Sigmod Rec, 1997, 26(1): 65-74.

[2] LO E, KAO B, HOW S, et al. OLAP on sequence data[C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. Vancouver: ACM, 2008: 649-660.

[3] CHUI C, KAO B, LO E, et al. S-olap: An olap system for analyzing sequence data[C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. Indianapolis: ACM, 2010: 1131-1134.

[4] MOHAMMAD S, SOUVIK B, BISHWARANJAN B, et al. L-Store: A Real-time OLTP and OLAP System[C]// Proceedings of the 21th International Conference on Extending Database Technology. Vienna: EDBT, 2018: 540-551.

[5] MOHAMADINA A A, GHAZALI M R B, IBRAHIM M R B, et al. Business intelligence: concepts issues and current systems [C]// 2012 International Conference on Advanced Computer Science Applications and Technologies. Kuala Lumpur, Malaysia: IEEE, 2012: 234-237.

[6] JOEL R, BRÁULIO A, SÉRGIO M. Business intelligence in a public institution — Evaluation of a financial data mart[C]// 12th Iberian Conference on Information Systems and Technologies (CISTI). Lisboa, Portugal, 2017: 1-6.

[7] GUO C P, WANG Z, HAN F, et al. HaoLap: An Hadoop Based OLAP System for Massive Data[J]. Journal of Computer Research and Development, 2013, 50(S1): 378-383.

[8] CHEN X Q, XU J, GUO J F, et al. BDA: An open big data analysis engine[J]. Newsletter of Chinese Computer Society, 2017, 13(8): 33-39.

[9] LI G J, CHEN X Q. Research Status and Scientific Thinking of Big Data[J]. Bulletin of the Chinese Academy of Sciences, 2012, 27(6): 647-651.

[10] PARK Y, MOZAFARI B, SORENSON J, et al. Verdictdb: Universalizing approximate query processing[C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: ACM, 2018: 1461-1476.

[11] WU Z G, JING Y N, HE Z Y, et al. POLYTOPE: a flexible sampling system for answering exploratory queries[J]. World Wide Web, 2019, 23(1): 1-22.

[12] CLEVELAND W S, MCGILL R. Graphical perception: Theory,

experimentation,and application to the development of graphical methods[J]. Journal of the American Statistical Association, 1984,79(387):531-554.

[13] MACKINLAY J D. Automating the design of graphical presentations of relational information[J]. ACM Transactions on Graphics,1986,5(2):110-141.

[14] LUO Y,QIN X,TANG N,et al. DeepEye:Towards Automatic Data Visualization[C]// 2018 IEEE 34th International Conference on Data Engineering. Paris:IEEE,2018:101-112.

[15] TROUILLON T,DANCE C R,GAUSSIER É,et al. Knowledge graph completion via complex tensor factorization[J]. The Journal of Machine Learning Research,2017,18(1):4735-4772.

[16] SUN Z Y,CHEN Z X,HE Z Y,et al. A Fast Automated Model Selection Approach Based on Collaborative Knowledge[J]. Database System for Advanced Applications,2020,12112:655-662.

[17] ASHISH T,JOYDEEP S S,NAMIT J,et al. Hive:a warehousing solution over a map-reduce framework[C]// Proceedings of the 35th International Conference on Very Large Data Bases. Lyon:Springer,2009:1626-1629.

YU Yue-zhang, born in 1997, master. His main research interest includes big data analysis.

JING Yi-nan, born in 1978, Ph.D, associate professor. His main research interests include big data analysis, spatial and temporal data management, mobile computing, and security and privacy.