面向最终用户的可定制数据统计服务①

刘祥龙[1,2], 许舒人[2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院软件研究所 软件工程技术研究开发中心, 北京 100190)

摘 要: 在传统企业应用中, 开发人员要根据不同的业务需求开发对应的数据统计模块, 这样的“预定义”数据统计方法很难满足不同用户对统计数据的个性化需求. 为了简化数据统计模块的开发流程, 同时满足不同用户的数据需求, 提出了面向最终用户的可定制数据统计服务解决方案. 本文着重介绍了面向用户的元数据模型和统计模型的表示方法, 以及基于动态构建 SQL 的即席查询方法, 并设计和实现了一套支持企业应用的数据统计服务.

关键词: 企业应用; 可定制; 数据统计; 即席查询

End-user Oriented Customizable Statistics Service

LIU Xiang-Long[1,2], XU Shu-Ren[2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Technology Center of Software Engineering, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: In traditional enterprise application, developers should develop different statistics modules according to corresponding business requirements. It is difficult to meet the individual needs of business users by using such “predefined” statistics methods. In order to simplify the development process of statistics modules and meet the individual data needs of users, this paper presents an end-user customizable statistics service solution. This paper focuses on the representation of user-oriented metadata model and statistical model, ad hoc query method based on dynamically constructed SQL, then presents the design and implement of a statistics service system supporting enterprise application.

Key words: enterprise application; customizable; metadata; Ad-hoc query

1 引言

随着 IT 技术的不断发展, 企业应用系统规模不断扩大, 功能模块不断增加, 开发和维护的成本也不断增加. 同时, 随着互联网技术的发展和硬件资源性能的提升使得企业获取数据和存储数据的门槛越来越低, 这个过程中产生的海量数据, 如何有效利用这些数据成为研究热点.

“云计算”为企业应用解决海量数据提供了新的解决方案, 它将不同业务线的功能模块和数据存储部署到统一的“云平台”环境下进行管理. 这样的做法将企业应用中的功能解耦出来打包成服务, 供客户端、移动端或其他系统调用, 极大地提高了系统的复用性和灵活性, 在提高效率的同时也便于进行特定功能数据分析工作的优化和升级. 在整合企业内部的计算资源和存储资源方面更加高效, 同时更加便于系统维护 [1].

在企业业务系统的众多功能中, 数据统计功能是与业务流程密切相关的基础功能, 以数据统计为支撑的报表、仪表板功能是每个业务系统中必不可少的部分. 过去我们通常在各个系统中嵌入与其业务相对应的数据统计模块, 实际上这些数据统计模块自身的运行逻辑大多数是一致的, 这也就在一定程度上造成重复的开发工作, 提高了系统开发成本. 同时, 传统企业应用中的数据统计通过预先定义的分析方法、报表、图表等为业务人员提供不同的数据获取方式. 这些预

① 基金项目:国家自然科学基金(61170074);国家科技支撑课题(2013BAH05F03);新闻出版重大科技工程项目(GXTC-CZ-1015004/01)
收稿时间:2015-03-28;收到修改稿时间:2015-04-26

定义的数据获取方式在系统开发阶段由技术人员设计实现, 由于他们与真正使用系统的决策者在业务知识上存在差异, 这些预先设计的数据获取方式并不一定能够完全满足决策者的业务需要, 这也就降低了数据统计功能的实用性.

在这样环境下, 建立一套基于服务的可定制数据统计服务, 能够从以下两点为企业应用提供帮助:

首先, 基于服务的实现方式将企业应用中各个业务系统将数据统计模块解耦出来, 形成数据统计服务. 这样的整合能够减少不同系统的重复开发工作, 同时解除了数据统计与各个系统的在实现技术上的依赖关系, 数据统计服务和业务系统都可以灵活地选择合适的技术进行功能实现. 这样的改变利于数据统计模块的快速迭代和升级, 也便于系统的扩展、维护.

第二, 统一的数据统计后台服务要能够提供更加灵活的接口服务于上层业务系统, 面对最终用户的需求变化, 不仅提供业务相关的预定义数据获取方式, 也可以开放给最终用户更多相关数据, 使其根据自己的需求定制数据, 进而定制适合自己的仪表板、报表等功能, 使得数据利用更加灵活有效.

因此, 本文通过分析数据统计功能特点和业务需求, 提出了相应的实现功能和满足需求的模型和方法, 并基于此设计和实现了一套面向最终用户的可定制数据统计服务, 以下将从问题分析、关键技术研究和系统设计与实现三个方面介绍.

2　问题分析

企业应用中, 商业智能系统作为业务决策支持系统是与数据接触最紧密的部分, 其根本任务是从业务过程产生的数据中获取更多的决策依据[2]. 随着业务数据不断增长, 用户需要更多样更灵活的数据统计方式来满足业务扩展和变化的需要.

文献[3]提出一种基于即席查询和业务协作的商业智能框架, 它将业务元数据从复杂的业务上下文中解耦出来, 提供更具灵活性的元数据模型, 使得业务人员能够在业务数据基础上通过可协作即席查询方法获取多样化的数据. 文献[4]则提出面向最终用户的“信息自服务”方法, 它通过整合不同用户场景的业务本体, 使得用户能够构建更具业务价值的语义查询方法, 使得其能够做出更好商业决策.

可以看出, 面向终端用户提供可定制的即席查询功能是满足多样性的用户需求的有效方法. 同时, 采用服务化方式构建数据统计系统, 避免了依赖于单一业务系统的实现方式, 为整个企业应用环境下的不同业务系统服务, 为企业应用部署方式带来更多灵活性和也能够带来更多性能上的优化.

基于这样的功能诉求, 本文介绍的数据统计服务作为企业应用的数据支撑, 通过获取业务数据库信息, 直接面向最终用户提供服务, 具体的服务方式如图 1 所示, 系统提供的服务功能分为三部分:

①元数据注册: 业务系统管理员将业务数据库的元数据注册到数据统计服务, 为统计服务访问业务数据库奠定基础.

②统计定制: 最终用户依据业务数据库元数据定制统计方式, 实现数据统计可定制.

③数据查询: 最终用户根据定制的数据统计方式, 查询业务数据获得统计图表.

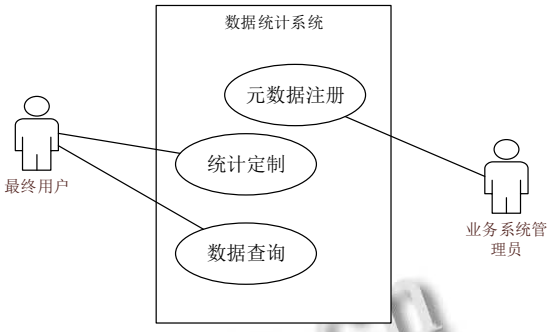

图 1　数据统计服务的服务方式

3　关键技术研究

为实现上述功能, 本节将介绍元数据模型、统计模型、即席查询三方面关键技术的研究.

3.1　元数据模型

数据统计服务的数据来源是业务系统的关系型数据库. 用户要操作这些数据, 需要系统在获取关系型数据库的元数据的同时也需要维护与元数据相对应的包含领域知识的语义数据, 使得元数据具备用户可读性. 文献[4]提出建立业务本体到技术本体的映射维护系统内部的元数据关系, 通过直接操作业务数据的方法为业务用户展示相关的业务视图. 而针对企业应用业务数据库而言, 对每一个领域对象都应在底层的关系型数据库上有对应的元数据(表、列、主键等)与之对应, 形成一个由语义元数据到数据库元数据的映射. 这样才能够保证底层的数据实现对上层业务系统是透

明的. 针对关系型数据库, 文献[5]提出的语义数据库模型是对数据库模型的更高层抽象, 它将数据库语义融合到数据库 Schema 中, 使得数据库设计者能够更加自然地与数据库进行交互. 我们利用语义数据库模型能够通过领域语义展现给最终用户能够获取到的数据的关系结构, 使其能够通过语义信息完成相应的配置和定制, 而不必过多关心底层的实现细节.

基于上述工作, 本系统使用的元数据模型除了维护与数据源相关的元数据外, 还要为数据统计功能密切相关的三种元数据: 表、属性、表连接, 配置一个领域词典. 领域字典中的每一个条目是由一个元数据到领域知识项的映射. 这样一个数据库的表结构就能够翻译得到含有业务语义的实体、属性、关系构成的实体关系图如图 2 所示, 变成用户可读的结构.

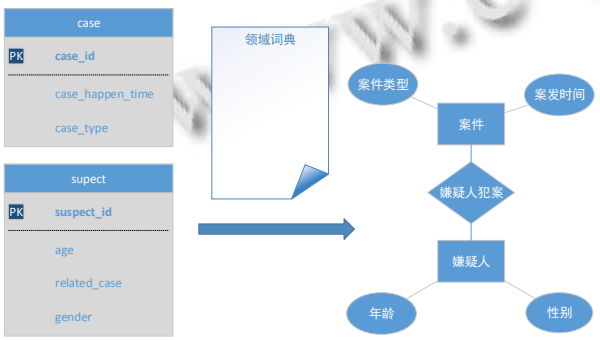

图 2 元数据模型转化

领域词典中的映射关系如表 1 所示:

表 1 领域词典的映射关系

语义元数据	数据库元数据	领域含义
实体	表	领域对象
属性	列	用于统计的属性
关系	表、列连接	领域对象的作用关系

例如, case 表映射为案件实体; case 表中的 case_happen_time 列映射为案发时间; suspect 表中 related_case 列是 case 表的外键, 将 case 表 case_id 列和 related_case 列映射为嫌疑人犯案关系.

3.2 统计模型

数据统计服务要建立面向最终用户的统计数据的定制机制. 这个定制过程需要一个形式化的统计模型来表达用户定制请求的数据结构, 而后数据统计服务通过统计模型得到用户指定的元数据信息, 最终获取指定的业务数据.

通常, 统计过程是按照某一个属性对另一个属性的规约过程, 例如, 各省案件数量对比统计, 是将案件的地案发地点对案件数量进行的规约. 同时, 在统计的过程中, 必要的约束条件也是保证数据统计满足业务需求的关键, 例如, 在进行各省案件数量统计时, 有时需要对时间范围进行限制, 需要添加 XX 年至 XX 年的案件数量对比.

文献[6]中分析了两种传统的统计模型的表示方法: 列表表示法和图表示法, 并且提出了一种基于图表示法的 STORM 模型. STORM 中首先将属性分为了统计量(能够进行数值计算)和分类量(无法进行数值计算), 而后将一个统计对象 SO 表示为一个四元组如公式 1:

$$SO = < N, C, S, F > \quad (1)$$

其中, N 表示统计对象名称, 描述了统计的领域内容; C 表示一个分类属性的有限集, 同时包含分类量的值域; S 表示一个与统计对象有关的统计属性, 同时包含统计属性的单位、统计类型; F 表示一个由分类属性的值的笛卡尔集到统计属性的值的映射函数.

STORM 模型完整地描述了定制一个数据统计所需要的所有语义信息, 然而在统计数据查询的场景中缺少了必要的条件约束, 例如统计某一个时间范围的量, 或是统计某一地区范围的量. 因此在定制统计对象需要添加相关的约束属性, 而约束属性具体的约束值则是在数据统计进行时即时动态绑定.

基于这样的需求, 本文将 STORM 模型扩展的五元组如公式 2:

$$\begin{aligned} SO &=< N,C,S,F,CONS > \\ CONS &= \{< Parameter, CON >\} \end{aligned} \quad (2)$$

添加的 $CONS$ 表示约束的有限集, 它用来指明统计参数限制的属性, 其中每一个约束有一个二元组组成, 其中 $Parameter$ 表示添加的约束参数; CON 表示与约束参数相关联的约束属性.

统计模型中约束参数是用户在进行查询时进行动态添加的, 其余的量决定了统计的形式. 根据这种不同我们将统计模型分为统计模板和统计参数, 前者包含除参数外的其余元素.

3.3 即席查询

即席查询是最终用户自己去建立特定的、自定义的查询请求[7]. 针对关系型数据库, 要完成数据统计的就要将用户请求转化成标准查询请求 SQL. 文献[8]研究了针对 MySQL 框架的 SQL 查询构建, 它使用语

言依赖和元数据构建可能 SELECT 和 WHERE 的子句集合, 同时挖掘元数据的连接关系构建 FROM 子句, 最终组成完整的 SQL 查询语句.

数据统计服务根据统计模型的元素所构建SQL语句进行即席查询, 除了 SELECT、WHERE、FROM 子句之外, 还会用的 GROUP BY 子句和相关的聚合函数. 所有子句中的信息都来自统计模型所维护的信息, 采用如图所示的 SQL 规则, 将统计模型转化而成的 SQL 语句包含 count、sum 等的聚合函数和 SELECT、FROM、WHERE、GROUP BY 子句.

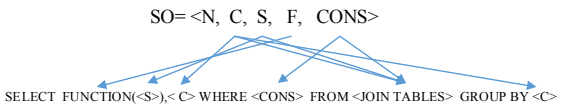

图 3　统计模型到 SQL 的转化

图 3 中统计模型中的元素分别对应到 SQL 的不同子句中: SELECT 子句包括统计属性、分类属性、映射函数; WHERE 子句包含约束属性; FROM 子句中的是由统计模型中所有涉及到的属性所属的表的连接得到; GROUP BY 子句包含分类属性.

可以看出在最终的 SQL 语句中只有 FROM 子句不能够从统计模型中的元素直接构成, 需要通过统计模型中设计到的所有数据库列, 找到各自相应的表后, 进行表连接. 在元数据模型的图结构(如图 2)中, 我们假设图中不存在环, 那么任意两个表之间的通过相关关系的连接可以得到一条路径, 计算这条路径通过图的搜索算法(算法 1 所示)得到, 将任意两表计算得到的路径上所有的表连接, 再去除重复连接, 即可得到表连接结构, 得到 FROM 子句.

系统将即时生成的 SQL 语句通过 JDBC 的执行即可得到需要的数据集.

算法 1　表连接算法

```
Input:entity1,entity2 //两个实体
Output: entity list//实体路径

FIND_CONNECTED_RELATION(e1, e2)
  var   paths //存放访问节点相邻实体
  var queue
  queue.offer(e1)
  while(queue is not empty)
     tmp = queue
     queue = new_queue()
     for(e in tmp)
        relations = getRelation(e)
        for(r in realtions)
          next = get_next_entity(r,e)//获取相邻实体
          if(paths.containsKey(next)) continue
          paths.put(next,r)//将实体加入路径中
          queue.offer(next);
          if(equals(next,e2))
             result = new_list();
             while(equals(next,entity1))
                r = paths.get(next);
                result.add(r);
                next = r.getOther(next);
             return result;
     return empty_list();//不存在实体连接
```

4　系统设计与实现

基于上文所提出的技术方法, 本文设计和实现了一套轻量级的数据统计服务系统, 提供元数据管理、统计模型管理、即席查询的三部分功能, 以下将从系统架构和功能实现详细进行介绍.

4.1 系统架构

本文所实现系统的系统架构如图 4 所示, 系统总体分为 Web 服务层、逻辑层、数据源层三部分, 下面详细介绍每个部分的作用

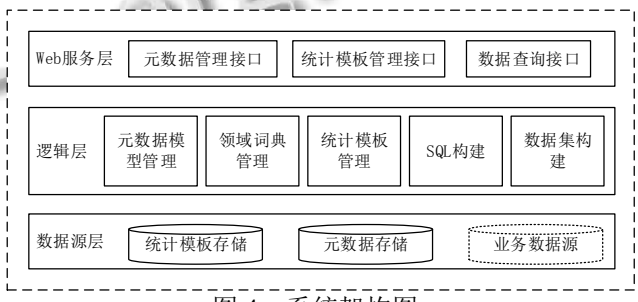

图 4　系统架构图

Web 服务层

Web 服务层采用 spring-boot 框架通过 HTTP 协议, 对外提供符合 REST[9]规范 API 接口, 使用 JSON 实现数据交换, 这样解耦了客户端和数据统计服务之间的依赖关系, 使得服务提供方式更加灵活.

Web 服务层接口包含元数据管理接口、统计模板管理、数据查询三部分接口, 为逻辑层完成元数据注

册、统计定制、和即席查询功能, 提供足够的数据交互接口.

逻辑层

逻辑层主要由五部分构成: 元数据模型管理、领域字典管理、统计模板管理、SQL 构建、数据集构建. 元数据管理和领域字典管理完成了从数据源的注册, 元数据的增删改查, 领域字典内容的增删改查. 统计模板管理根据用户输入和元数据模型完成统计模板的增删改查. SQL 构建和数据集构建将用户的定制信息转换成 SQL, 收集 SQL 执行得到的数据进行处理后返回给用户.

数据源层

提供逻辑层数据的持久化机制, 其中元数据存储是将元数据模型持久化到数据库; 统计模板存储是将用户定制的统计模板持久化到数据库; 业务数据库(虚线表示)是业务系统的关系型数据存储, 系统对其只有读取权限.

4.2 功能实现

4.2.1 元数据注册

元数据注册功能的实现如图 5 所示, 共分为三个步骤:

第一步, 数据源注册. 本系统通过 jdbc 与业务数据库建立连接, 这个过程除了要求业务数据库开放数据统计服务的访问权限外, 还需要用户提供数据库类型、业务数据库 IP 地址、端口号、数据库名称等相关信息. 这些信息通过数据源注册生成一个 Database 实例, Database 中的数据库连接信息保证系统在之后的运行过程中能够获取业务数据库相关信息.

第二步, 数据库元数据注册. 在这一步用户输入与数据统计相关的数据库表名、数据库列名, 系统通过 Database 与业务数据库建立的连接获取与表、列相关的元数据分别生成 Table、Column 实例.

第三步, 领域知识注册. 在系统已经注册的数据库元数据(表数据、列数据)基础上, 用户输入对应的元数据领域描述, 得到在领域概念上的实体 Entity 和属性 Attribute, 并且将表之间的有连接关系的列转化为关系.

在完成上述三个步骤后, 系统即完成了对一个业务数据库的元数据注册过程. 我们将元数据注册过程中产生的所有元数据组织到元数据模型类 Schema 中, Schema 的类图如图 6 所示. 每一个业务数据库注册完成后生成的 Schema 实例, 系统为其分配唯一的 UUID, 以区分不同数据库的元数据模型.

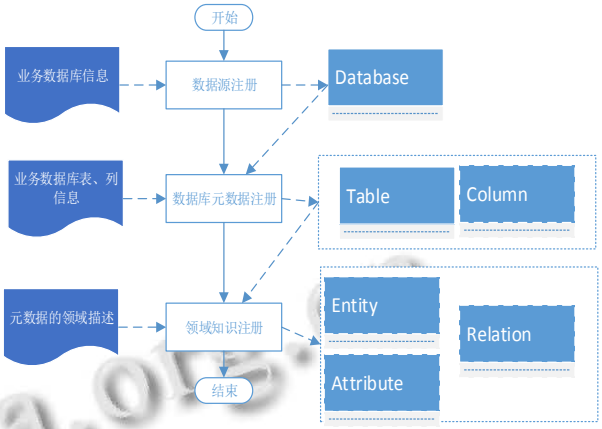

图 5 元数据注册流程图

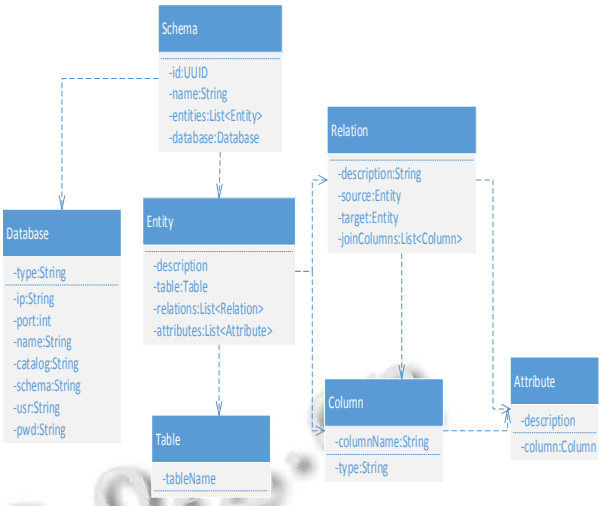

图 6 元数据模型类类图

4.2.2 统计模板定制

数据统计的定制是基于统计模型的元素需要, 用户对不同属性的选择过程. 根据公式(2)、(3)所表示的统计模型的内容, 统计模型中 CONS 约束中的参数是用户在查询进行动态添加的, 在用户定制统计方式时并不存在, 根据这种不同, 系统将统计模型分为统计模板(类 Template)和统计参数(类 Parameter)两部分. 对数据系统的定制实际上是对定制一个统计模板实例, 因此, 如图 7 所示, 用户需要首先输入统计的描述信息, 然后选择统计属性 summary、分类属性集 categories、映射方法 function、约束属性集 constraints, 完成所有的选择过程后, 系统就可以生成统计模板 Template 实例.

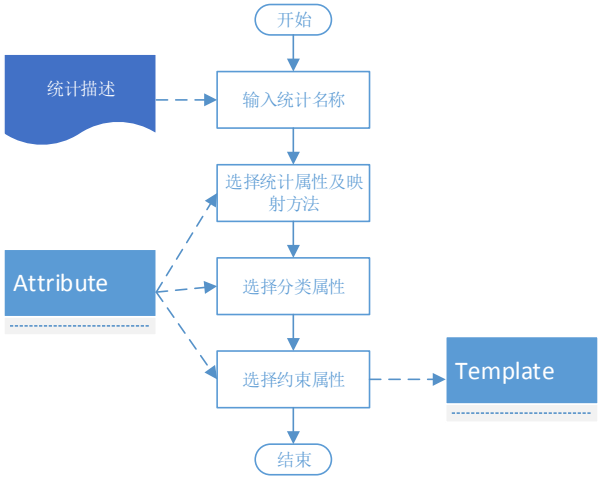

图 7　统计模板定制流程图

4.2.3 数据查询

数据查询过程是用户根据定制的统计模板动态生成的查询，数据查询过程如图 8 所示.

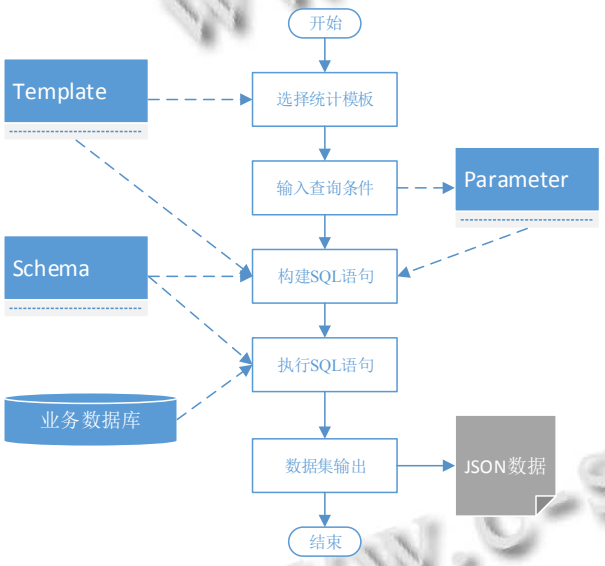

图 8　即席查询过程

用户首先选择某一个统计模板实例，根据这个模板实例中定义的约束属性类型，输入与每个约束属性相对应的约束参数 Parameter. 这样统计模板 Template 和统计参数 Parameter 就能够组成一个完整的统计模型. 根据图 3 所示的 SQL 构建规则以及算法 1 所示的表连接方法，将统计模型转化为相对应的 SQL 语句. 构建好的 SQL 语句与业务数据库建立的 jdbc 连接执行 SQL 语句后得到了结果集 ResultSet. 最后，将 ResultSet 数据整理输出得到 JSON 格式的数据集返回给用户，这样就完成了一次统计查询.

综上所述系统实现了由元数据注册、统计模板定制、数据查询的功能，能够为用户提供完整的可定制的数据统计服务.

4.3 系统验证

本文实现的系统 Sostats 已稳定运行在某《犯罪信息系统》中，本节以某《犯罪信息系统》(下称: 信息系统)为例验证本文的系统 Sostats 实现的功能.

4.3.1 元数据注册

信息系统使用 Oracle 11g 数据库存储业务数据，在系统启动时，首先将数据源信息注册到 Sostats 服务中. 这里的数据源信息保存在配置文件中，如代码 1 所示，信息系统启动后自动读取该文件信息通过 JAX-WS 注册到 Sostats 系统中.

代码 1 配置文件 Config.properties 内容

```
url=localhost
port=1521
username=sostats
password=test
database=cccis
type=ORACLE
schema=SYSTEM
```

接下来，将信息系统的元数据注册到系统中. 信息系统中的数据库元数据以及对应的领域描述信息保存在配置文件 StatisticsContext.xml 中，部分内容如代码 2 所示，其中 context 标签为 xml 根标签; table 标签保存了表信息; column 标签保存了列信息; table-join 标签保存了关系信息. 信息系统在注册数据源后，读取 StatisticsContext.xml 中的元数据信息，同样通过 JAX-WS 接口注册到 Sostats 系统中. 这样，信息系统就完成了元数据注册.

代码 2 StatisticalContext.xml 文件部分内容

```
<context>
    ...
    <table name="case_info" description="案件信息"
distinctBy="id">
        <column type="category" name="case_location"
description="案发地点" />
        <column type="summary" name="seized_amount"
description="公安收缴量" />
        <column type="condition" name="register_date"
description="立案时间" />
        ...
```

```
        </table>
    <table-join description="银行记录">
        <joined table="jiabi_base_info" column="fmid"/>
        <joining table="jiabi_from_bank" column="fmid"/>
    </table-join>
    …
</context>
```

4.3.2 统计模版定制

信息系统启动后, 通过 Sostats 的接口获取元数据模型的信息, 进行统计模板的定制, 界面如图 9 所示. 用户在定制统计模板时, 需要填写: 统计图名称、统计属性、统计函数、分类属性以及约束属性. 变量的填写通过选择下拉表中的领域描述完成. 用户点击“保存”后, 创建的模板显示在统计分析的列表中, 如图 10 所示.

图 9 统计模板定制界面

图 10 统计分析模板列表

4.3.3 数据查询

用户选择统计分析列表中的模板进入对应统计模板的查询界面, 如图 11 所示. 界面中显示了相关的条件输入框, 这与统计模板中的约束属性的选择对应. 输入条件后, Sostats 进行处理得到 JSON 格式的结果集后, 通过 echarts 数据可视化库的渲染得到相应的结果集图, 如图 12 所示, 这样用户就完成了完整的数据统计流程.

图 11 数据查询条件输入界面

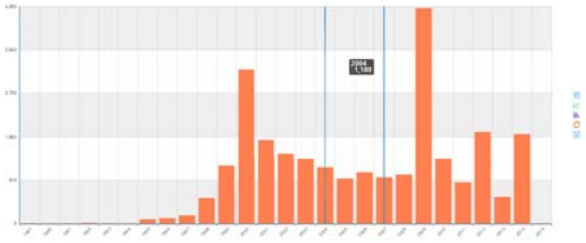

图 12 数据查询结果

5 总结

本文基于面向最终用户定制化的思想提出企业应用系统中常用的数据统计服务的新方案, 首先分析了数据统计服务所要解决的问题, 实现了元数据模型、统计模型和基于动态 SQL 生成的即席查询的关键技术. 在这些技术方法的基础上, 设计和实现了一套面向最终用户的可定制数据统计服务系统. 系统能够提供业务用户在不同业务场景下统计分析相关仪表板、统计图、报表的所需要的后台数据的定制功能, 同时简化了数据服务的开发, 能够帮助提高系统开发的效率.

参考文献

1 Ouf S, Nasr M. The cloud computing: the future of BI in the cloud. International Journal of Computer Theory and Engineering, 2011, 3(6): 750–754

2 Chaudhuri S, Dayal U, Narasayya V. An overview of business intelligence technology. Communications of the ACM, 2011, 54(8): 88–98.

3 Berthold H, Rösch P, Zöller S, et al. An architecture for ad-hoc and collaborative business intelligence. Proc. of the 2010 EDBT/ICDT Workshops. ACM, 2010: 13.

4 Spahn M, Kleb J, Grimm S, et al. Supporting business intelligence by providing ontology-based end-user information self-service. Proc. of the First international Workshop on ontology-Supported Business intelligence. ACM, 2008: 10.

5 Hammer M, Mc Leod D. Database description with SDM: a semantic database mode. ACM Trans. on Database Systems (TODS), 1981, 6(3): 351–386.

6 Rafanelli M, Shoshani A. Storm: A statistical object representation model. Springer Berlin Heidelberg, 1990.

7 Ad-hoc, http://zh.wikipedia.org/wiki/Ad_hoc

8 Giordani A, Moschitti A. Generating SQL queries using natural language syntactic dependencies and metadata. Natural Language Processing and Information Systems. Springer Berlin Heidelberg, 2012: 164–170.

9 Fielding R. Representational state transfer. Architectural Styles and the Design of Netowork-based Software Architecture, 2000: 76–85.